LECTURE 6

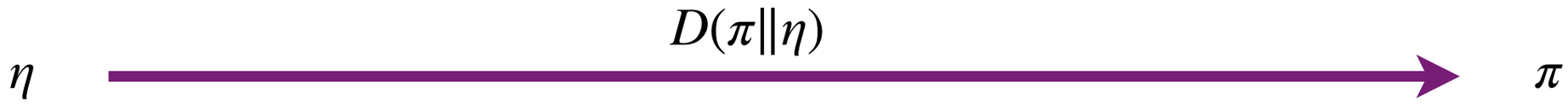
ANNEALING

Saifuddin Syed

ANNEALING

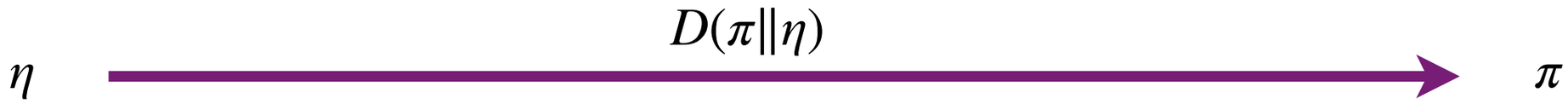
ANNEALING

- Given a divergence D and a reference η and target π



ANNEALING

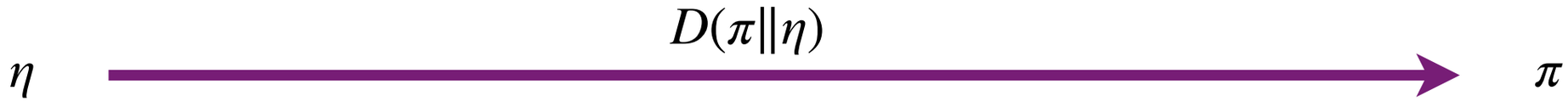
- Given a divergence D and a reference η and target π



- **Solution 2:** Keep the reference fixed and modify the target

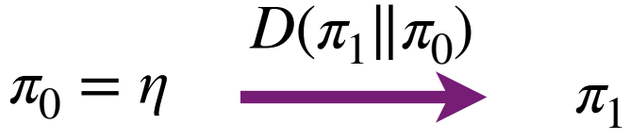
ANNEALING

- Given a divergence D and a reference η and target π



- **Solution 2:** Keep the reference fixed and modify the target

  - We can effieciently propagate inferences from $\eta = \pi_0$ to π_1 if $D(\pi_1\|\pi_0)$ is small



ANNEALING

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{D(\pi\|\eta)} \pi$$

- **Solution 2:** Keep the reference fixed and modify the target

  - We can effieciently propagate inferences from $\eta = \pi_0$ to π_1 if $D(\pi_1\|\pi_0)$ is small

$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1$$

  - We can then efficiently propogate inference from π_1 to π_2 close to π_1

$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1 \xrightarrow{D(\pi_2\|\pi_1)} \pi_2$$

ANNEALING

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{D(\pi\|\eta)} \pi$$

- **Solution 2:** Keep the reference fixed and modify the target

  - We can effieciently propagate inferences from $\eta = \pi_0$ to π_1 if $D(\pi_1\|\pi_0)$ is small

$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1$$

  - We can then efficiently propogate inference from π_1 to π_2 close to π_1

$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1 \xrightarrow{D(\pi_2\|\pi_1)} \pi_2$$

  - Can repeat until we reach the target

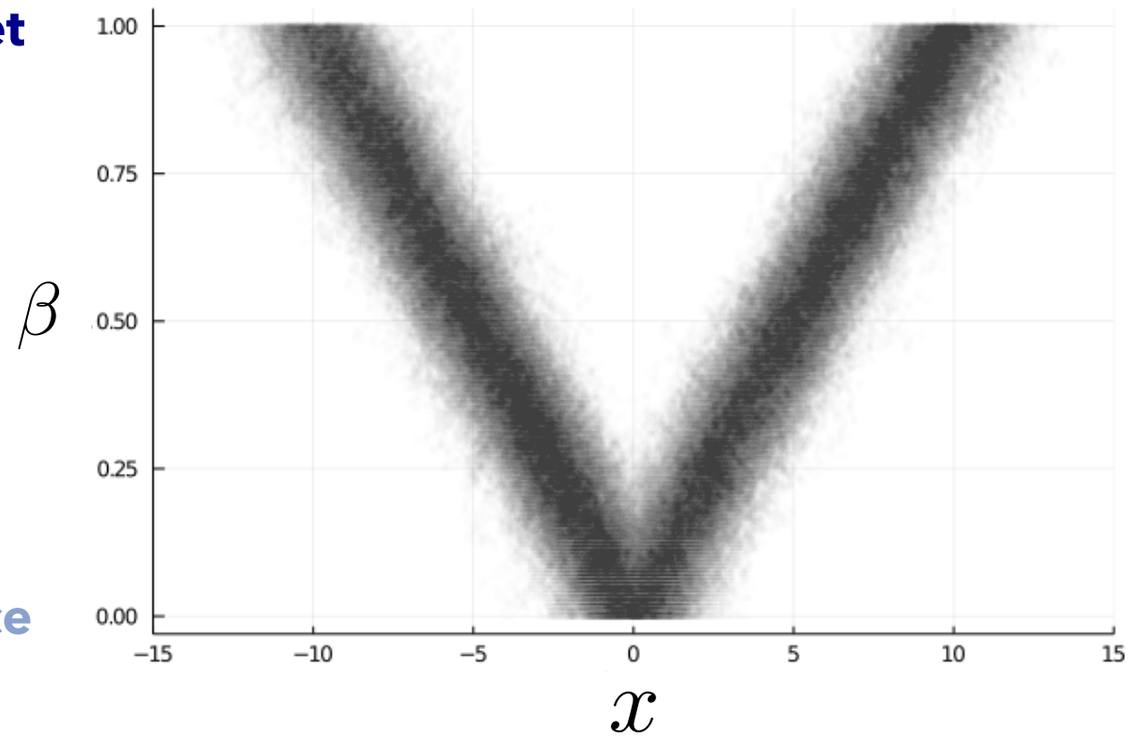
$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1 \xrightarrow{D(\pi_2\|\pi_1)} \pi_2 \quad \cdots \quad \pi_{N-1} \xrightarrow{D(\pi_N\|\pi_{N-1})} \pi_N = \pi$$

ANNEALING

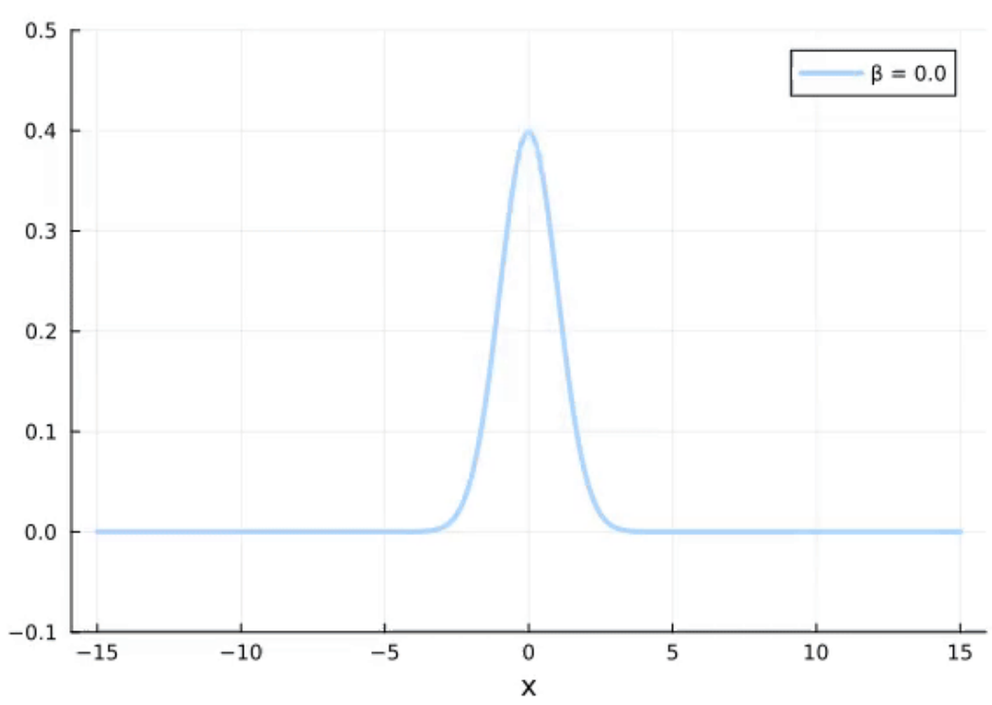
ANNEALING

- For $\beta \in [0,1]$ suppose we hav distributions π_β such that $\pi_0 = \eta$ and $\pi_1 = \pi$
  - π_β is the **annealing distribution** corresponding to the **annealing parameter** β

Target



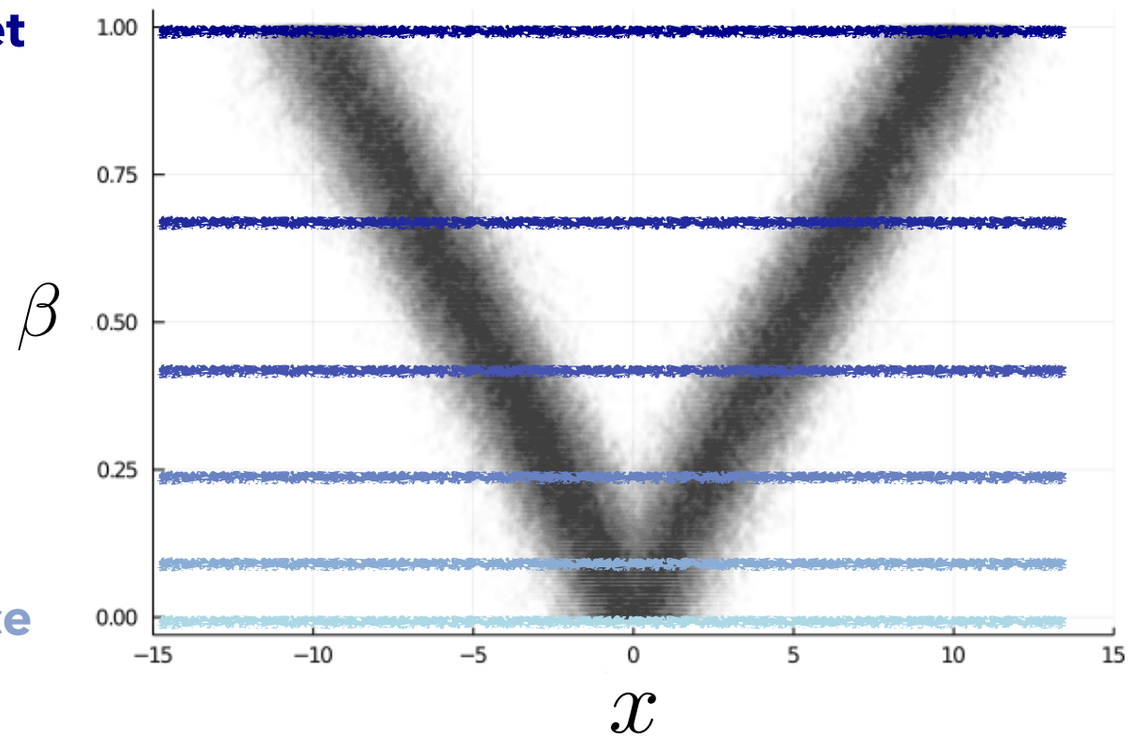
Reference



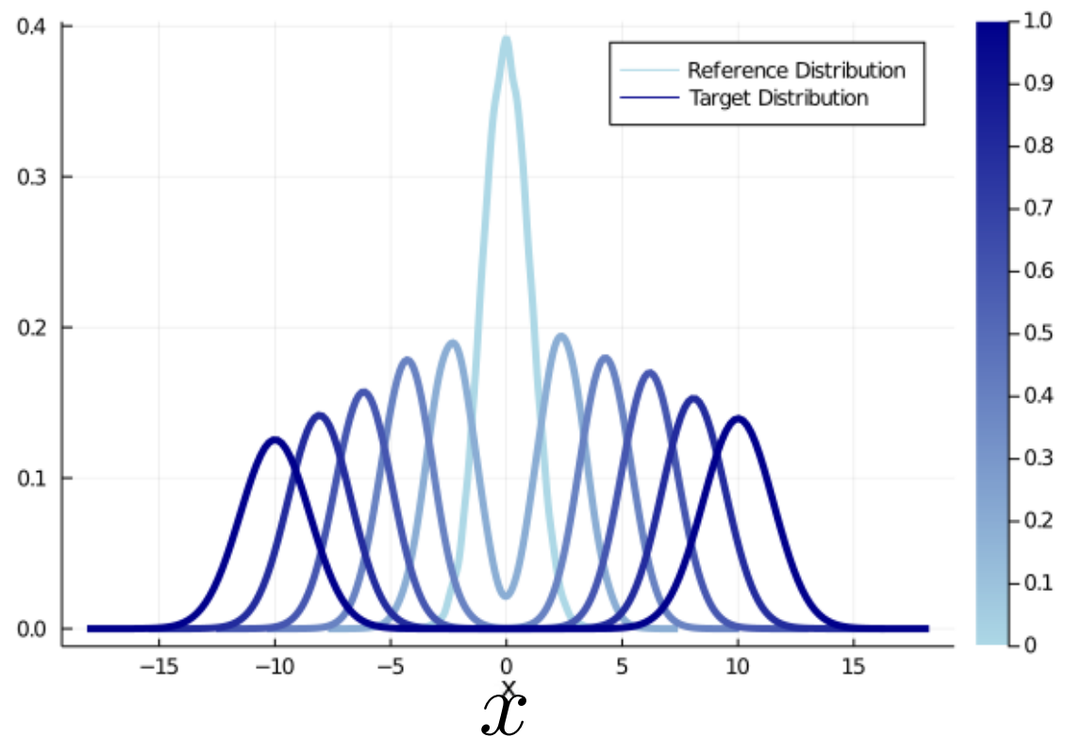
ANNEALING

- For $\beta \in [0,1]$ suppose we hav distributions π_β such that $\pi_0 = \eta$ and $\pi_1 = \pi$
  - π_β is the **annealing distribution** corresponding to the **annealing parameter** β

Target



Reference



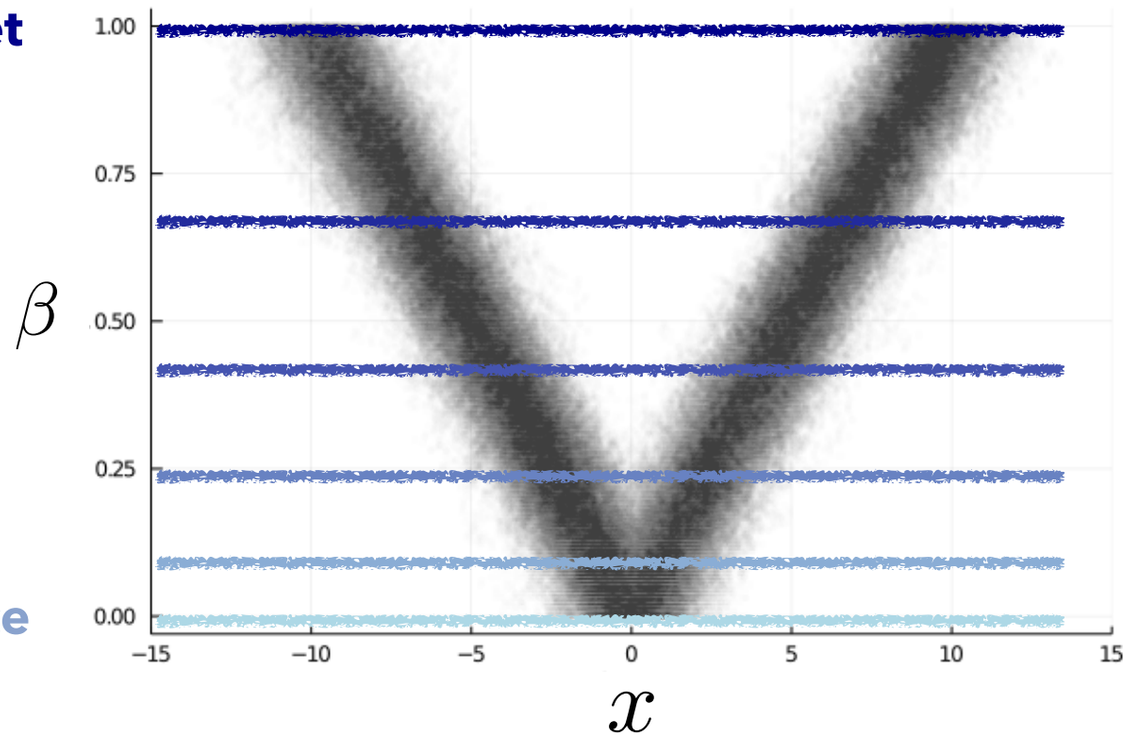
ANNEALING

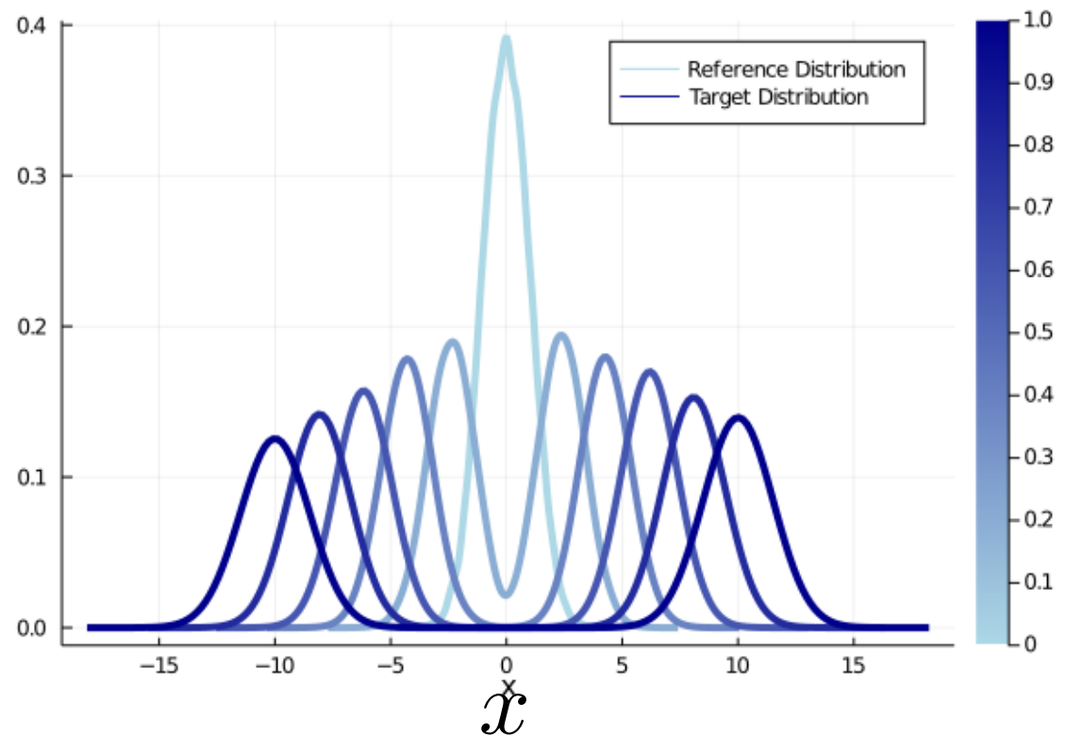
- For $\beta \in [0,1]$ suppose we hav distributions π_β such that $\pi_0 = \eta$ and $\pi_1 = \pi$
  - π_β is the **annealing distribution** corresponding to the **annealing parameter** β
  - Where $\mathscr{B} = \beta_{0:N}$ is the **annealing schedule** satisfying:

$$0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1$$

Target

Reference





ANNEALING

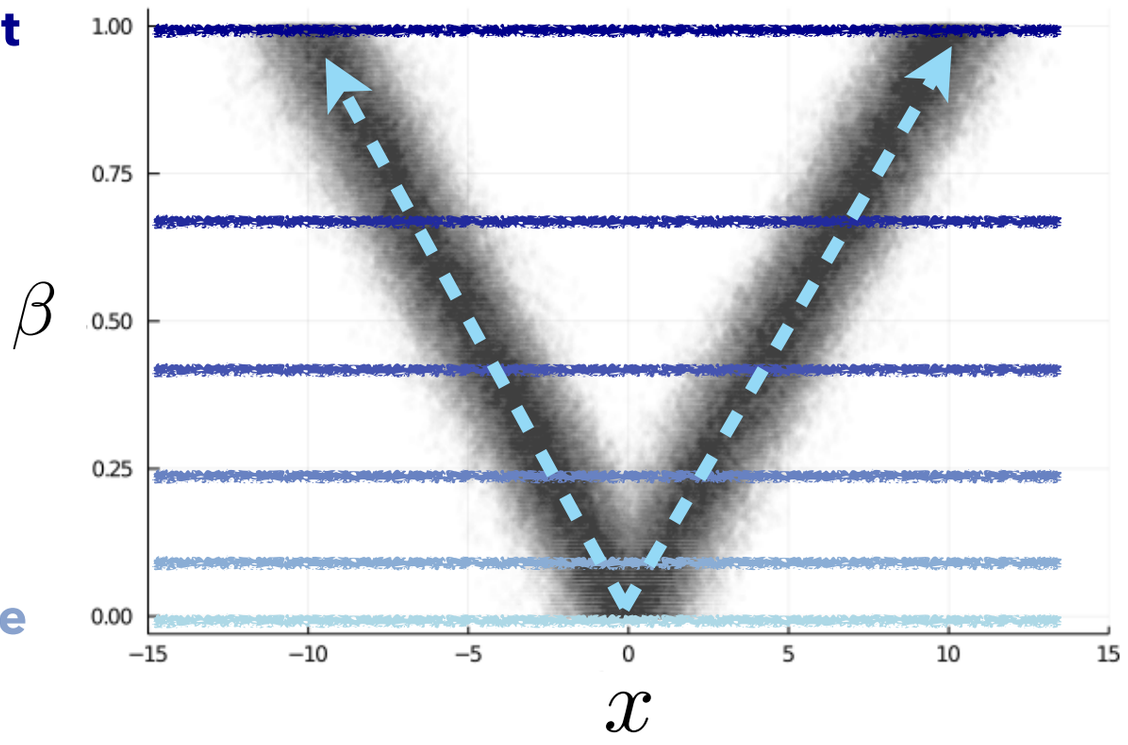
- For $\beta \in [0,1]$ suppose we hav distributions π_β such that $\pi_0 = \eta$ and $\pi_1 = \pi$
  - π_β is the **annealing distribution** corresponding to the **annealing parameter** β
  - Where $\mathscr{B} = \beta_{0:N}$ is the **annealing schedule** satisfying:

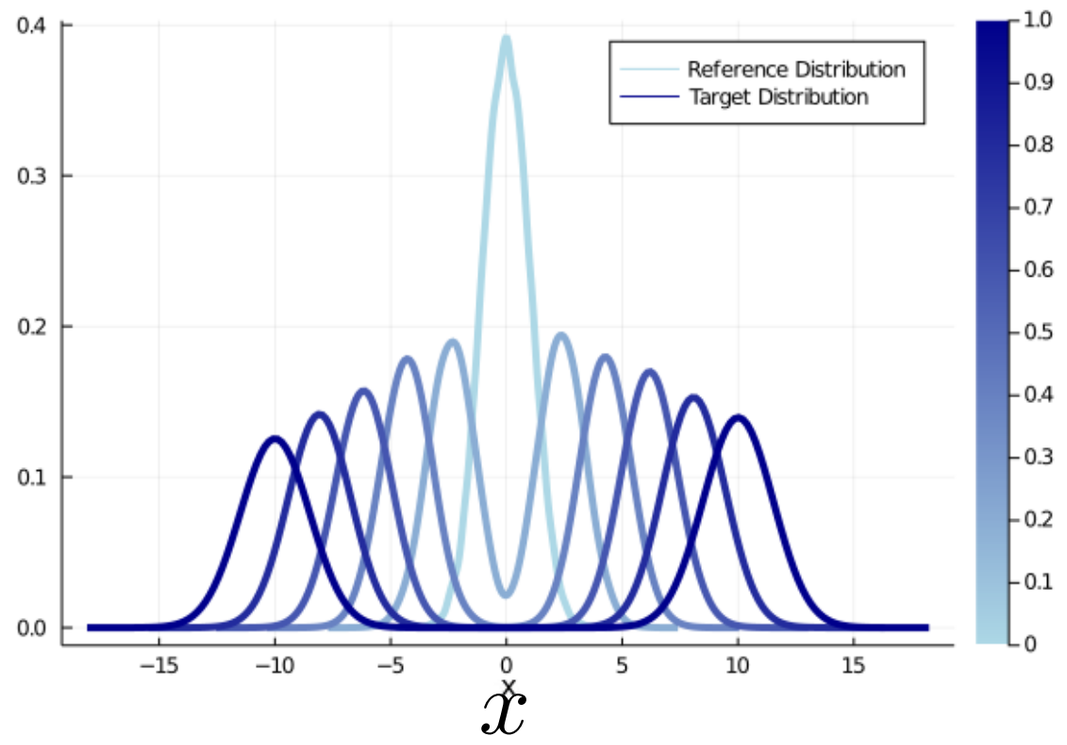
$$0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1$$

- Transform a d-dimensional multi-modal target into a $d+1$-dimensional unimodal one

Target

Reference





HOW TO CONSTRUCT AN ANNEALING PATH

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η

$$\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x, \qquad \pi(\mathrm{d}x) = \pi(x)\mathrm{d}x = \frac{\gamma(x)}{Z}\mathrm{d}x$$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η

$$\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x, \qquad \pi(\mathrm{d}x) = \pi(x)\mathrm{d}x = \frac{\gamma(x)}{Z}\mathrm{d}x$$

- We build annealing distribution by curupting the densities:

$$\pi_\beta(x) = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η

$$\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x, \qquad \pi(\mathrm{d}x) = \pi(x)\mathrm{d}x = \frac{\gamma(x)}{Z}\mathrm{d}x$$

- We build annealing distribution by curupting the densities:

$$\pi_\beta(x) = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

  - $\beta \mapsto \gamma_\beta(x)$ is a continuous function for all $x \in \mathbb{X}$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η

$$\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x, \qquad \pi(\mathrm{d}x) = \pi(x)\mathrm{d}x = \frac{\gamma(x)}{Z}\mathrm{d}x$$

- We build annealing distribution by curupting the densities:

$$\pi_\beta(x) = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

  - $\beta \mapsto \gamma_\beta(x)$ is a continuous function for all $x \in \mathbb{X}$
  - $\gamma_0 = \eta$ and $\gamma_1 = \gamma$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η

$$\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x, \qquad \pi(\mathrm{d}x) = \pi(x)\mathrm{d}x = \frac{\gamma(x)}{Z}\mathrm{d}x$$

- We build annealing distribution by curupting the densities:

$$\pi_\beta(x) = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

  - $\beta \mapsto \gamma_\beta(x)$ is a continuous function for all $x \in \mathbb{X}$
  - $\gamma_0 = \eta$ and $\gamma_1 = \gamma$
  - $Z(\beta) < \infty$ for all $\beta \in [0,1]$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η

$$\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x, \qquad \pi(\mathrm{d}x) = \pi(x)\mathrm{d}x = \frac{\gamma(x)}{Z}\mathrm{d}x$$

- We build annealing distribution by curupting the densities:

$$\pi_\beta(x) = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

  - $\beta \mapsto \gamma_\beta(x)$ is a continuous function for all $x \in \mathbb{X}$
  - $\gamma_0 = \eta$ and $\gamma_1 = \gamma$
  - $Z(\beta) < \infty$ for all $\beta \in [0,1]$

- We build annealing distribution by curupting the densities:

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η

$$\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x, \qquad \pi(\mathrm{d}x) = \pi(x)\mathrm{d}x = \frac{\gamma(x)}{Z}\mathrm{d}x$$

- We build annealing distribution by curupting the densities:

$$\pi_\beta(x) = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

  - $\beta \mapsto \gamma_\beta(x)$ is a continuous function for all $x \in \mathbb{X}$
  - $\gamma_0 = \eta$ and $\gamma_1 = \gamma$
  - $Z(\beta) < \infty$ for all $\beta \in [0,1]$

- We build annealing distribution by curupting the densities:

$$V_\beta = \log \gamma_\beta, \qquad A(\beta) = \log Z(\beta)$$

LINEAR PATH

LINEAR PATH

- The canonical choice is the linear/geometric path

$$\pi_\beta \propto \eta^{1-\beta}\pi^\beta$$

LINEAR PATH

- The canonical choice is the linear/geometric path

$$\pi_\beta \propto \eta^{1-\beta}\pi^\beta$$

  - Has un-normalised density:

$$\begin{aligned}\gamma_\beta(x) &= \eta(x)^{1-\beta}\gamma(x)^\beta \\ &= \eta(x)w(x)^\beta, \qquad w(x) = \frac{\gamma(x)}{\eta(x)}\end{aligned}$$

LINEAR PATH

- The canonical choice is the linear/geometric path

$$\pi_\beta \propto \eta^{1-\beta}\pi^\beta$$

  - Has un-normalised density:

$$\begin{aligned}\gamma_\beta(x) &= \eta(x)^{1-\beta}\gamma(x)^\beta \\ &= \eta(x)w(x)^\beta, \qquad w(x) = \frac{\gamma(x)}{\eta(x)}\end{aligned}$$

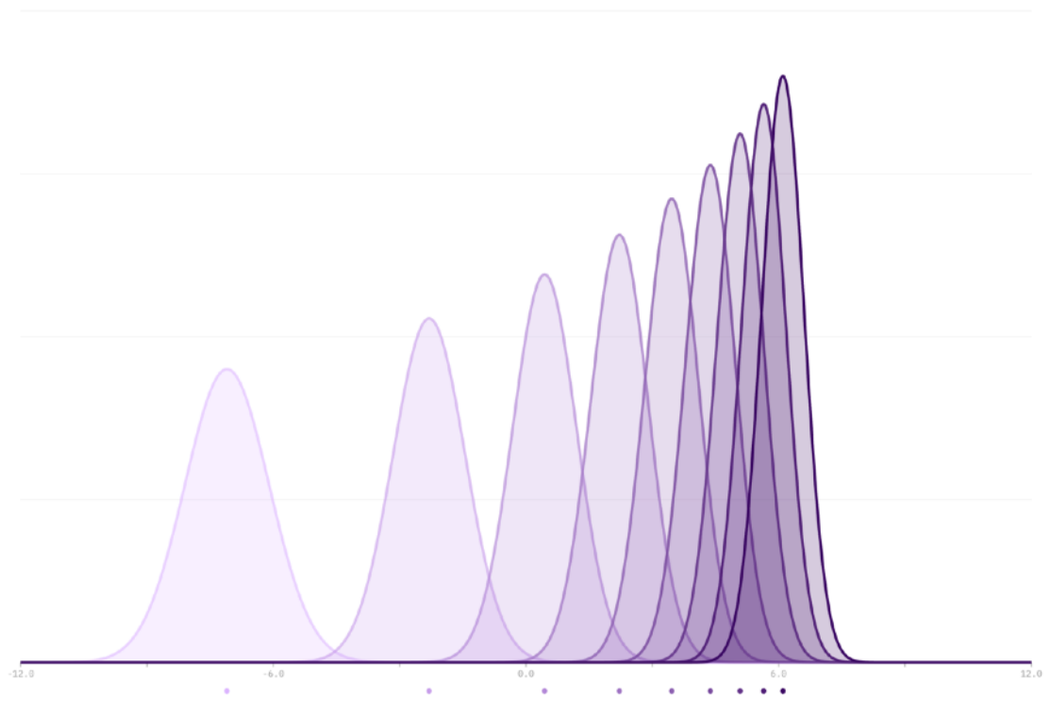
  - Equivalently, it corresponds to the linear interpolation between the log-densities

$$\begin{aligned}V_\beta &= (1-\beta)V_0 + \beta V_1 \\ &= V_0 + \beta V, \qquad V = \log w\end{aligned}$$

- The linear path flattens the likelihood ratio between referece and target

$$\frac{\mathrm{d}\pi_\beta}{\mathrm{d}\eta}(x) \propto \left(\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x)\right)^\beta \propto w(x)^\beta$$

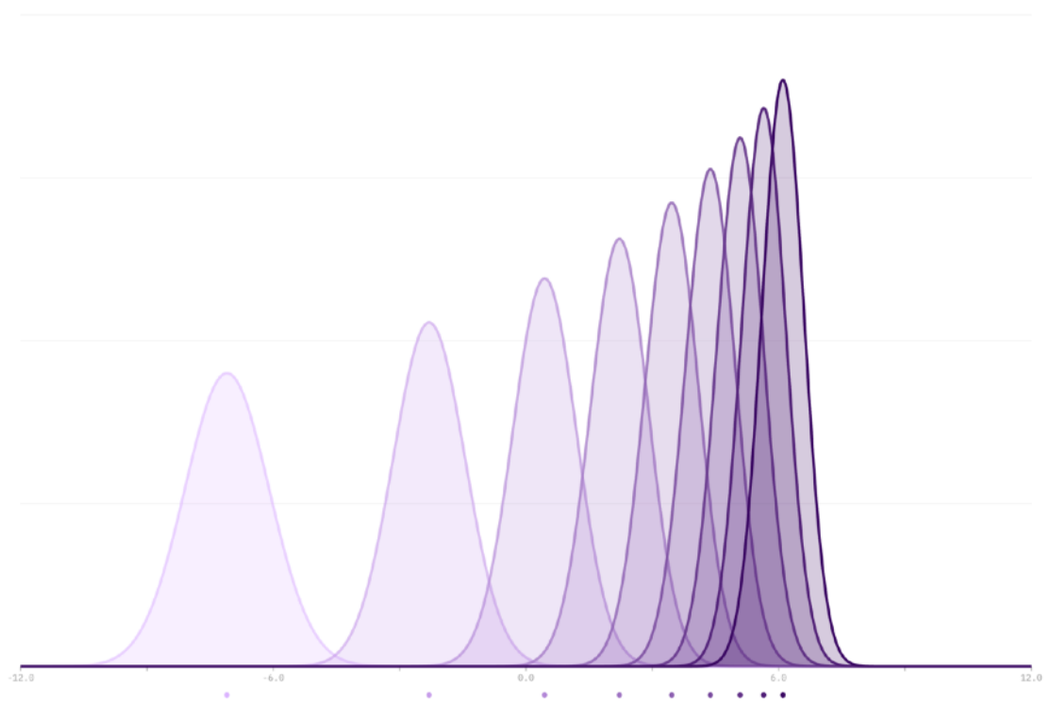
LINEAR PATH



LINEAR PATH

- This object is independent of the state-space $\mathbb{X}$, making it very generally applicable

$$\pi_\beta \propto \eta^{1-\beta}\pi^\beta$$

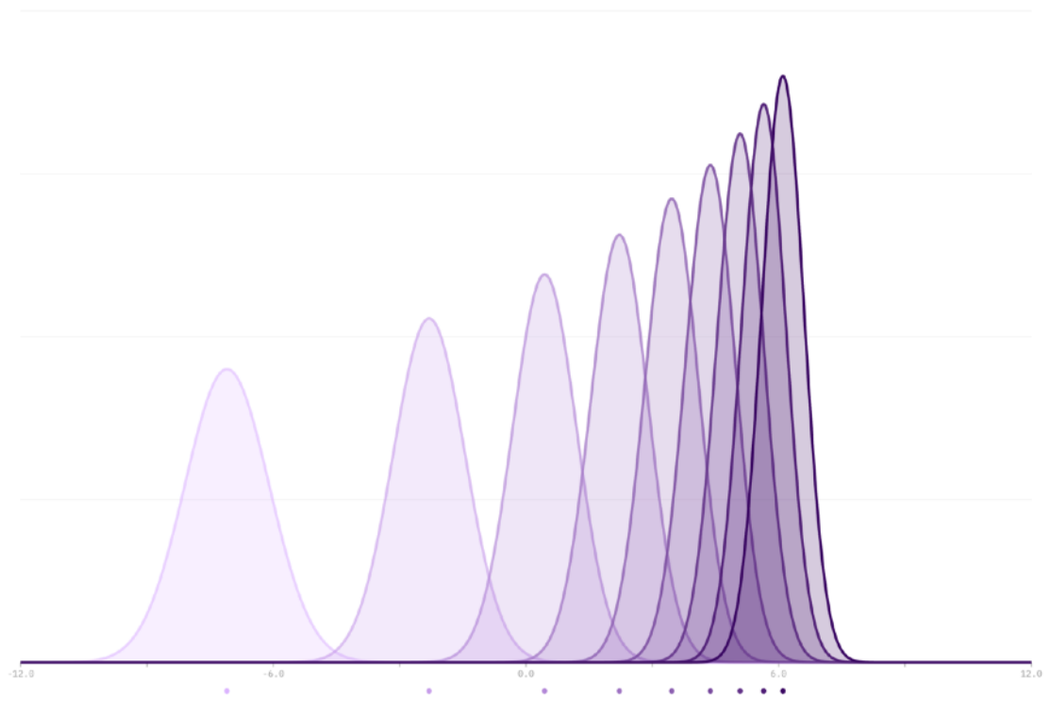


LINEAR PATH

- This object is independent of the state-space $\mathbb{X}$, making it very generally applicable

$$\pi_\beta \propto \eta^{1-\beta}\pi^\beta$$

  - It has been independently been discovered and applied in a variety of domains

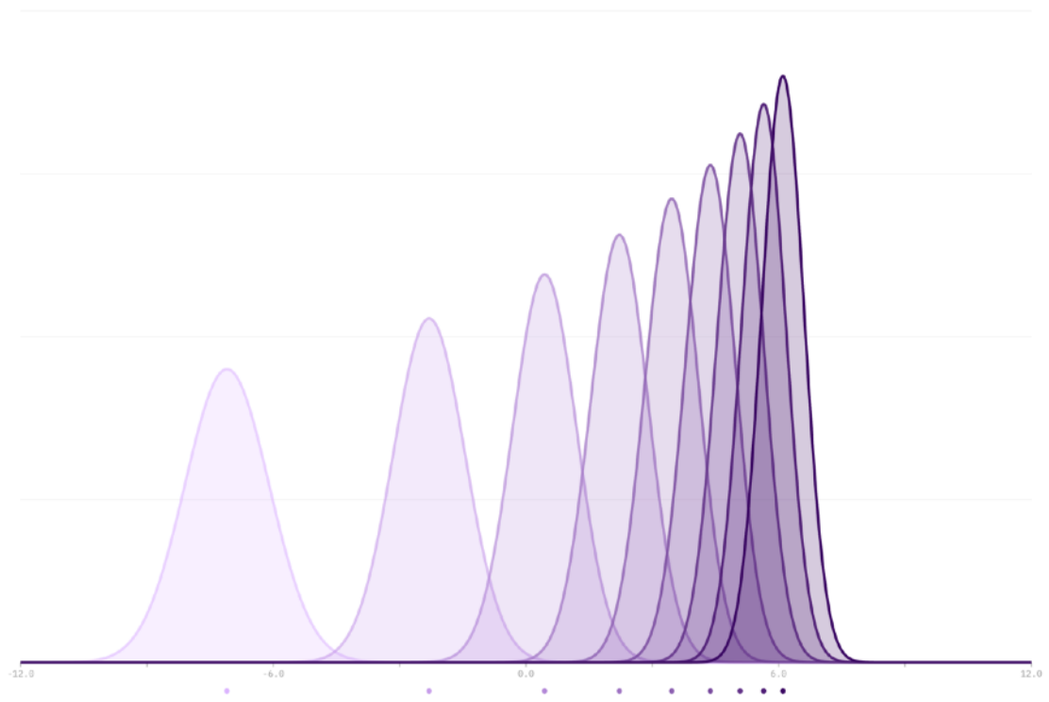


LINEAR PATH

- This object is independent of the state-space $\mathbb{X}$, making it very generally applicable

$$\pi_\beta \propto \eta^{1-\beta}\pi^\beta$$

  - It has been independently been discovered and applied in a variety of domains
  - e.g. probability, Bayesian, information geomerty, optimisation, and statistical mechanics



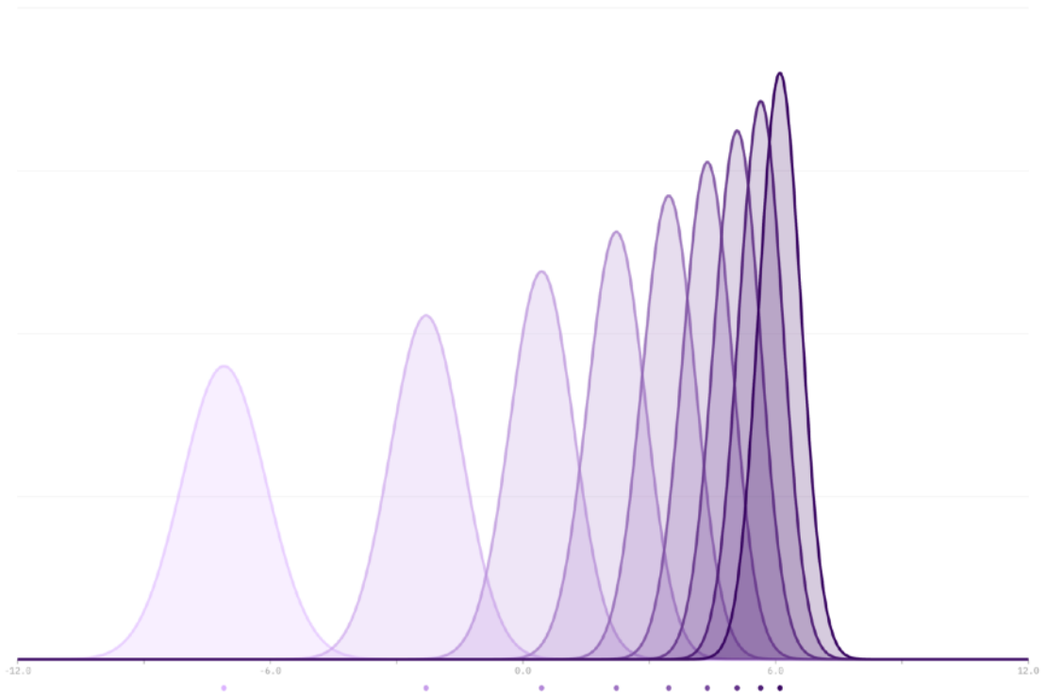
LINEAR PATH

- This object is independent of the state-space $\mathbb{X}$, making it very generally applicable

$$\pi_\beta \propto \eta^{1-\beta}\pi^\beta$$

  - It has been independently been discovered and applied in a variety of domains
  - e.g. probability, Bayesian, information geomerty, optimisation, and statistical mechanics

- The linear path greedily matches densities point wise for each x



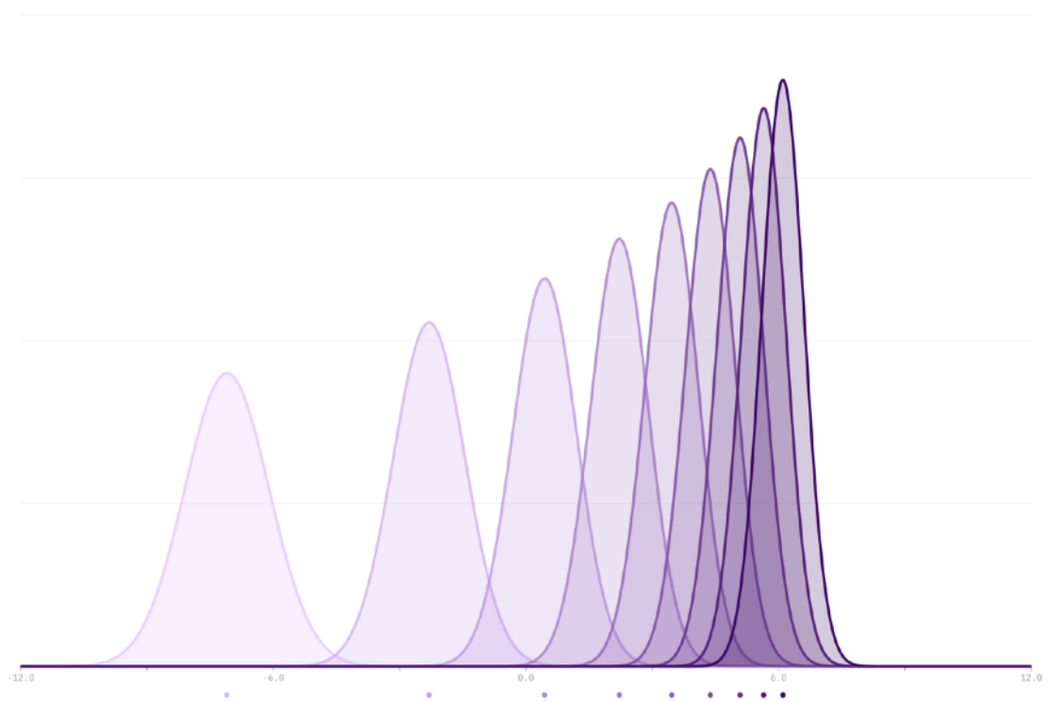
LINEAR PATH

- This object is independent of the state-space $\mathbb{X}$, making it very generally applicable

$$\pi_\beta \propto \eta^{1-\beta}\pi^\beta$$

  - It has been independently been discovered and applied in a variety of domains
  - e.g. probability, Bayesian, information geomerty, optimisation, and statistical mechanics

- The linear path greedily matches densities point wise for each x
  - It is rarely optimal for sampling, but is practical



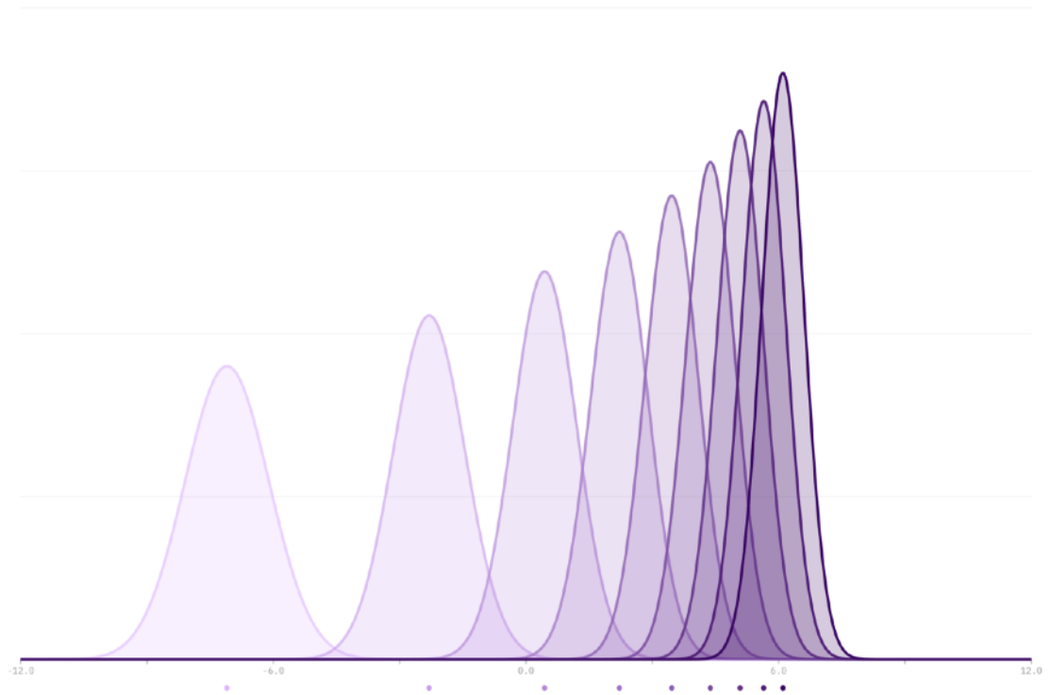
LINEAR PATH

- This object is independent of the state-space $\mathbb{X}$, making it very generally applicable

$$\pi_\beta \propto \eta^{1-\beta}\pi^\beta$$

  - It has been independently been discovered and applied in a variety of domains
  - e.g. probability, Bayesian, information geomerty, optimisation, and statistical mechanics

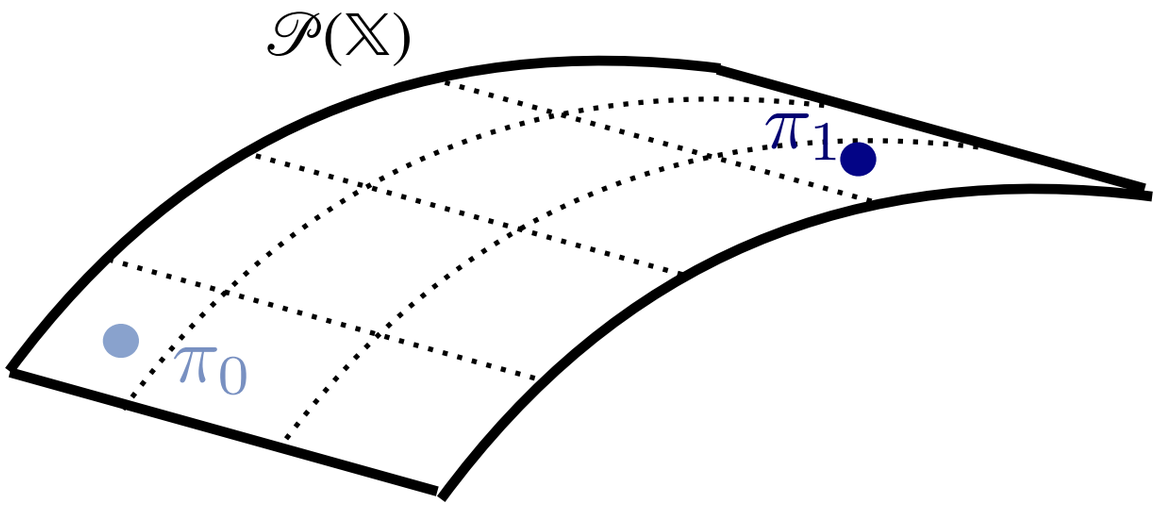
- The linear path greedily matches densities point wise for each x
  - It is rarely optimal for sampling, but is practical
  - We will see later how to design better paths



ANNEALING AS A CURVE

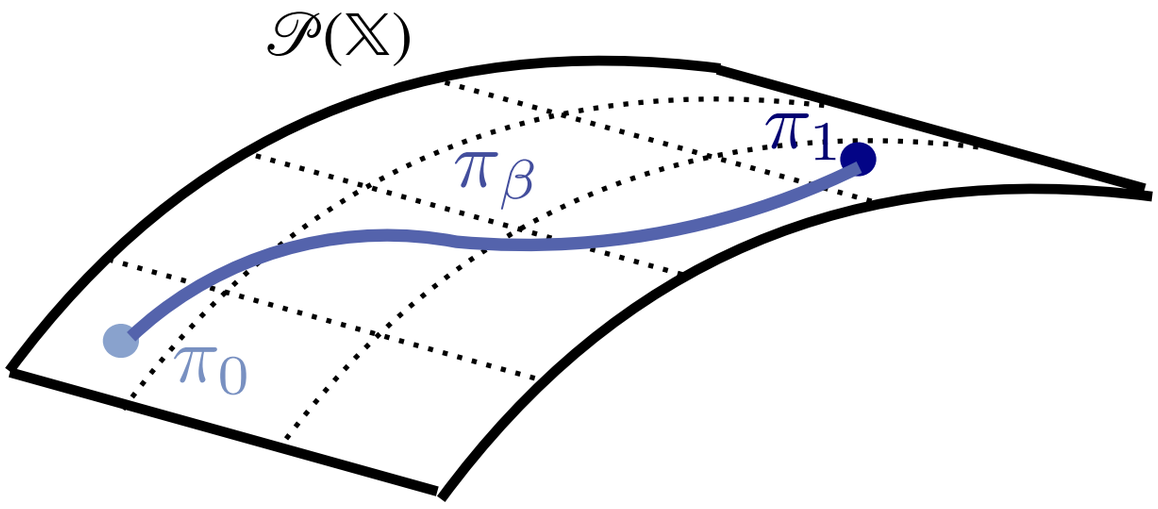
ANNEALING AS A CURVE

- Suppose $\mathscr{P}(\mathbb{X})$ is the space of probability densities supported on $\mathbb{X}$



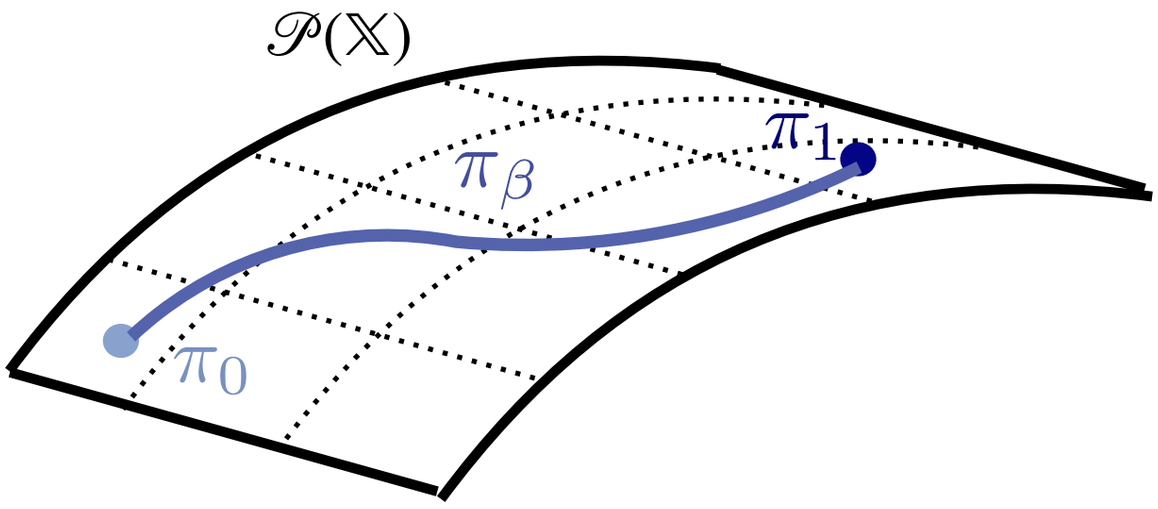
ANNEALING AS A CURVE

- Suppose $\mathscr{P}(\mathbb{X})$ is the space of probability densities supported on $\mathbb{X}$
- $\beta \mapsto \pi_\beta$ defines a curve in $\mathscr{P}(\mathbb{X})$ parametrised by β with **position** π_β



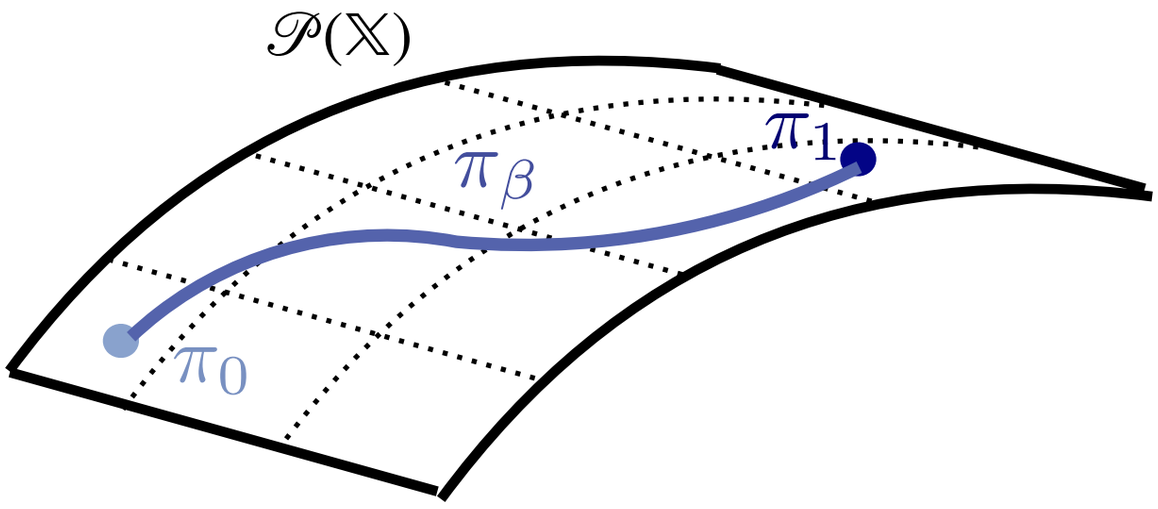
ANNEALING AS A CURVE

- Suppose $\mathscr{P}(\mathbb{X})$ is the space of probability densities supported on $\mathbb{X}$
- $\beta \mapsto \pi_\beta$ defines a curve in $\mathscr{P}(\mathbb{X})$ parametrised by β with **position** π_β
- We can measure how much the distribution changes from β to β' through the likelihood ratio



ANNEALING AS A CURVE

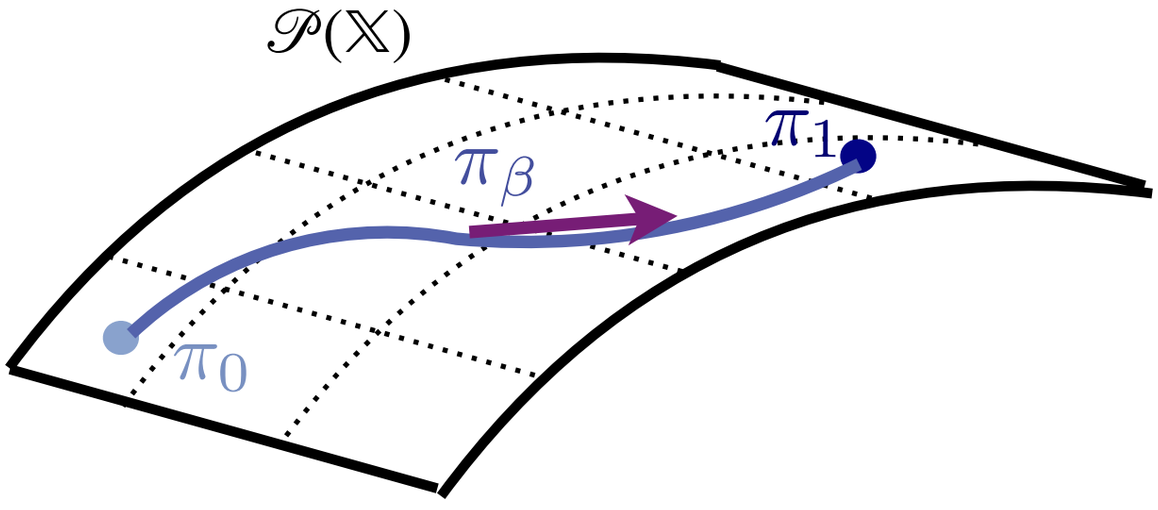
- Suppose $\mathscr{P}(\mathbb{X})$ is the space of probability densities supported on $\mathbb{X}$
- $\beta \mapsto \pi_\beta$ defines a curve in $\mathscr{P}(\mathbb{X})$ parametrised by β with **position** π_β
- We can measure how much the distribution changes from β to β' through the likelihood ratio
- The change in position $\Delta\pi_{\beta,\beta'} : \mathbb{X} \mapsto \mathbb{R}_+$ as the percentage change in likelihood

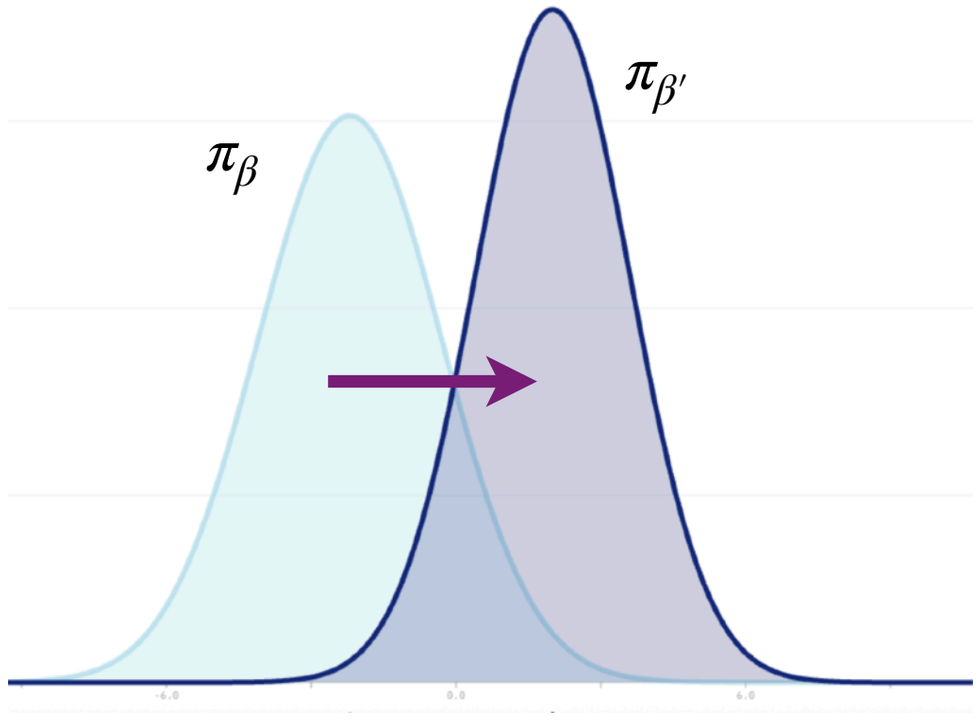


ANNEALING AS A CURVE

- Suppose $\mathscr{P}(\mathbb{X})$ is the space of probability densities supported on $\mathbb{X}$
- $\beta \mapsto \pi_\beta$ defines a curve in $\mathscr{P}(\mathbb{X})$ parametrised by β with **position** π_β
- We can measure how much the distribution changes from β to β' through the likelihood ratio
- The change in position $\Delta\pi_{\beta,\beta'} : \mathbb{X} \mapsto \mathbb{R}_+$ as the percentage change in likelihood

$$\Delta\pi_{\beta,\beta'}(x) = \frac{\mathrm{d}\pi_{\beta'}}{\mathrm{d}\pi_\beta}(x) - 1 = \frac{\pi_{\beta'}(x) - \pi_\beta(x)}{\pi_\beta(x)}$$





VELOCITY OF DENSITY

VELOCITY OF DENSITY

- Define the **velocity** of $\dot{\pi}_\beta : \mathbb{X} \to \mathbb{R}$ as the instantaneous rate position

$$\dot{\pi}_\beta(x) := \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta}$$

VELOCITY OF DENSITY

- Define the **velocity** of $\dot{\pi}_\beta : \mathbb{X} \to \mathbb{R}$ as the instantaneous rate position

$$\dot{\pi}_\beta(x) := \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta}$$

- The velocity coincides with the **Fisher score** encodes the direction and rate of change in density

$$\dot{\pi}_\beta(x) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

VELOCITY OF DENSITY

- Define the **velocity** of $\dot{\pi}_\beta : \mathbb{X} \to \mathbb{R}$ as the instantaneous rate position

$$\dot{\pi}_\beta(x) := \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta}$$

- The velocity coincides with the **Fisher score** encodes the direction and rate of change in density

$$\dot{\pi}_\beta(x) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

- Note that for the linear path, the Fisher score constant velocity (upto an additive constant)

$$\dot{\pi}_\beta(x) = V(x) - \frac{\mathrm{d}}{\mathrm{d}\beta} A(\beta)$$

VELOCITY OF DENSITY

- Define the **velocity** of $\dot{\pi}_\beta : \mathbb{X} \to \mathbb{R}$ as the instantaneous rate position

$$\dot{\pi}_\beta(x) := \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta}$$

- The velocity coincides with the **Fisher score** encodes the direction and rate of change in density

$$\dot{\pi}_\beta(x) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

  - Note that for the linear path, the Fisher score constant velocity (upto an additive constant)

$$\dot{\pi}_\beta(x) = V(x) - \frac{\mathrm{d}}{\mathrm{d}\beta} A(\beta)$$

- **Proof:**

VELOCITY OF DENSITY

- Define the **velocity** of $\dot{\pi}_\beta : \mathbb{X} \to \mathbb{R}$ as the instantaneous rate position

$$\dot{\pi}_\beta(x) := \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta}$$

- The velocity coincides with the **Fisher score** encodes the direction and rate of change in density

$$\dot{\pi}_\beta(x) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

  - Note that for the linear path, the Fisher score constant velocity (upto an additive constant)

$$\dot{\pi}_\beta(x) = V(x) - \frac{\mathrm{d}}{\mathrm{d}\beta} A(\beta)$$

- **Proof:**

$$\dot{\pi}_\beta(x) = \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta}$$

VELOCITY OF DENSITY

- Define the **velocity** of $\dot{\pi}_\beta : \mathbb{X} \to \mathbb{R}$ as the instantaneous rate position

$$\dot{\pi}_\beta(x) := \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta}$$

- The velocity coincides with the **Fisher score** encodes the direction and rate of change in density

$$\dot{\pi}_\beta(x) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

  - Note that for the linear path, the Fisher score constant velocity (upto an additive constant)

$$\dot{\pi}_\beta(x) = V(x) - \frac{\mathrm{d}}{\mathrm{d}\beta} A(\beta)$$

- **Proof:**

$$\dot{\pi}_\beta(x) = \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta} = \lim_{\Delta\beta \to 0} \frac{\pi_{\beta+\Delta\beta}(x) - \pi_\beta(x)}{\Delta\beta\pi_\beta(x)}$$

VELOCITY OF DENSITY

- Define the **velocity** of $\dot{\pi}_\beta : \mathbb{X} \to \mathbb{R}$ as the instantaneous rate position

$$\dot{\pi}_\beta(x) := \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta}$$

- The velocity coincides with the **Fisher score** encodes the direction and rate of change in density

$$\dot{\pi}_\beta(x) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

  - Note that for the linear path, the Fisher score constant velocity (upto an additive constant)

$$\dot{\pi}_\beta(x) = V(x) - \frac{\mathrm{d}}{\mathrm{d}\beta} A(\beta)$$

- **Proof:**

$$\dot{\pi}_\beta(x) = \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta} = \lim_{\Delta\beta \to 0} \frac{\pi_{\beta+\Delta\beta}(x) - \pi_\beta(x)}{\Delta\beta\pi_\beta(x)} = \frac{1}{\pi_\beta(x)} \frac{\mathrm{d}\pi_\beta}{\mathrm{d}\beta}(x)$$

VELOCITY OF DENSITY

- Define the **velocity** of $\dot{\pi}_\beta : \mathbb{X} \to \mathbb{R}$ as the instantaneous rate position

$$\dot{\pi}_\beta(x) := \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta}$$

- The velocity coincides with the **Fisher score** encodes the direction and rate of change in density

$$\dot{\pi}_\beta(x) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

  - Note that for the linear path, the Fisher score constant velocity (upto an additive constant)

$$\dot{\pi}_\beta(x) = V(x) - \frac{\mathrm{d}}{\mathrm{d}\beta} A(\beta)$$

- **Proof:**

$$\dot{\pi}_\beta(x) = \lim_{\Delta\beta \to 0} \frac{\Delta\pi_{\beta,\beta+\Delta\beta}(x)}{\Delta\beta} = \lim_{\Delta\beta \to 0} \frac{\pi_{\beta+\Delta\beta}(x) - \pi_\beta(x)}{\Delta\beta\pi_\beta(x)} = \frac{1}{\pi_\beta(x)} \frac{\mathrm{d}\pi_\beta}{\mathrm{d}\beta}(x) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

VELOCITY OF EXPECTATIONS

VELOCITY OF EXPECTATIONS

- For $f : \mathbb{X} \to \mathbb{R}$ the velocity $\dot{\boldsymbol{\pi}}_\beta$ measures how the expectation $\boldsymbol{\pi}_\beta[f]$ change with

VELOCITY OF EXPECTATIONS

- For $f : \mathbb{X} \to \mathbb{R}$ the velocity $\dot{\pi}_\beta$ measures how the expectation $\pi_\beta[f]$ change with

- **Proposition:** Assume π_β and f are sufficiently regular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] = \pi_\beta[f\dot{\pi}_\beta]$$

VELOCITY OF EXPECTATIONS

- For $f : \mathbb{X} \to \mathbb{R}$ the velocity $\dot{\pi}_\beta$ measures how the expectation $\pi_\beta[f]$ change with

- **Proposition:** Assume π_β and f are sufficiently regular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] = \pi_\beta[f\dot{\pi}_\beta]$$

- **Proof:** Assume π_β and f are sufficiently regular, we have

VELOCITY OF EXPECTATIONS

- For $f : \mathbb{X} \to \mathbb{R}$ the velocity $\dot{\pi}_\beta$ measures how the expectation $\pi_\beta[f]$ change with

- **Proposition:** Assume π_β and f are sufficiently regular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] = \pi_\beta[f\dot{\pi}_\beta]$$

- **Proof:** Assume π_β and f are sufficiently regular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] = \frac{\mathrm{d}}{\mathrm{d}\beta}\int_{\mathbb{X}} f(x)\pi_\beta(x)\mathrm{d}x$$

VELOCITY OF EXPECTATIONS

- For $f : \mathbb{X} \to \mathbb{R}$ the velocity $\dot{\pi}_\beta$ measures how the expectation $\pi_\beta[f]$ change with

- **Proposition:** Assume π_β and f are sufficiently regular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] = \pi_\beta[f\dot{\pi}_\beta]$$

- **Proof:** Assume π_β and f are sufficiently regular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] = \frac{\mathrm{d}}{\mathrm{d}\beta}\int_{\mathbb{X}} f(x)\pi_\beta(x)\mathrm{d}x$$

$$= \int_{\mathbb{X}} f(x)\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta(x)\mathrm{d}x$$

VELOCITY OF EXPECTATIONS

- For $f : \mathbb{X} \to \mathbb{R}$ the velocity $\dot{\pi}_\beta$ measures how the expectation $\pi_\beta[f]$ change with

- **Proposition:** Assume π_β and f are sufficiently regular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] = \pi_\beta[f\dot{\pi}_\beta]$$

- **Proof:** Assume π_β and f are sufficiently regular, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] &= \frac{\mathrm{d}}{\mathrm{d}\beta}\int_{\mathbb{X}} f(x)\pi_\beta(x)\mathrm{d}x \\ &= \int_{\mathbb{X}} f(x)\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta(x)\mathrm{d}x \\ &= \int_{\mathbb{X}} f(x)\dot{\pi}_\beta(x)\pi_\beta(x)\mathrm{d}x\end{aligned}$$

VELOCITY OF EXPECTATIONS

- For $f : \mathbb{X} \to \mathbb{R}$ the velocity $\dot{\pi}_\beta$ measures how the expectation $\pi_\beta[f]$ change with

- **Proposition:** Assume π_β and f are sufficiently regular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] = \pi_\beta[f\dot{\pi}_\beta]$$

- **Proof:** Assume π_β and f are sufficiently regular, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f] &= \frac{\mathrm{d}}{\mathrm{d}\beta}\int_{\mathbb{X}} f(x)\pi_\beta(x)\mathrm{d}x \\
&= \int_{\mathbb{X}} f(x)\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta(x)\mathrm{d}x \\
&= \int_{\mathbb{X}} f(x)\dot{\pi}_\beta(x)\pi_\beta(x)\mathrm{d}x \\
&= \pi_\beta[f\dot{\pi}_\beta]
\end{aligned}$$

VELCOCITY OF ANNEALING PATHS

VELCOCITY OF ANNEALING PATHS

- In summary, we can equivalently express the velocity $\dot{\boldsymbol{\pi}}_\beta$ terms of densities and expectations

VELCOCITY OF ANNEALING PATHS

- In summary, we can equivalently express the velocity $\dot{\pi}_\beta$ terms of densities and expectations

- **Velocity as a function:** $x \mapsto \dot{\pi}_\beta(x)$ defines a function equals to the instantaneous percentage change in density $\pi_\beta(x)$ at β for all $x \in \mathbb{X}$

$$\dot{\pi}_\beta(x) := \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

VELCOCITY OF ANNEALING PATHS

- In summary, we can equivalently express the velocity $\dot{\pi}_\beta$ terms of densities and expectations

- **Velocity as a function:** $x \mapsto \dot{\pi}_\beta(x)$ defines a function equals to the instantaneous percentage change in density $\pi_\beta(x)$ at β for all $x \in \mathbb{X}$

$$\dot{\pi}_\beta(x) := \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

- **Velocity as a measure:** $f \mapsto \dot{\pi}_\beta[f]$ defines a (signed)-measure equal to the instantaneous rate of change in expectation $\pi_\beta[f]$ a β for $f : \mathbb{X} \mapsto \mathbb{R}$ with $f(x)\dot{\pi}_\beta(x)$ is integrable with respect to π_β:

$$\dot{\pi}_\beta[f] := \frac{\mathrm{d}}{\mathrm{d}\beta} \pi_\beta[f]$$

VELCOCITY OF ANNEALING PATHS

- In summary, we can equivalently express the velocity $\dot{\pi}_\beta$ terms of densities and expectations

- **Velocity as a function:** $x \mapsto \dot{\pi}_\beta(x)$ defines a function equals to the instantaneous percentage change in density $\pi_\beta(x)$ at β for all $x \in \mathbb{X}$

$$\dot{\pi}_\beta(x) := \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

- **Velocity as a measure:** $f \mapsto \dot{\pi}_\beta[f]$ defines a (signed)-measure equal to the instantaneous rate of change in expectation $\pi_\beta[f]$ a β for $f : \mathbb{X} \mapsto \mathbb{R}$ with $f(x)\dot{\pi}_\beta(x)$ is integrable with respect to π_β:

$$\dot{\pi}_\beta[f] := \frac{\mathrm{d}}{\mathrm{d}\beta} \pi_\beta[f]$$

- We will abuse notation and will identify π_β with both the density or distributional derivative

VELCOCITY OF ANNEALING PATHS

- In summary, we can equivalently express the velocity $\dot{\pi}_\beta$ terms of densities and expectations

- **Velocity as a function:** $x \mapsto \dot{\pi}_\beta(x)$ defines a function equals to the instantaneous percentage change in density $\pi_\beta(x)$ at β for all $x \in \mathbb{X}$

$$\dot{\pi}_\beta(x) := \frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

- **Velocity as a measure:** $f \mapsto \dot{\pi}_\beta[f]$ defines a (signed)-measure equal to the instantaneous rate of change in expectation $\pi_\beta[f]$ a β for $f : \mathbb{X} \mapsto \mathbb{R}$ with $f(x)\dot{\pi}_\beta(x)$ is integrable with respect to π_β:

$$\dot{\pi}_\beta[f] := \frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f]$$

- We will abuse notation and will identify π_β with both the density or distributional derivative

- We can relate the derivatives of the expectation and densities through the identity:

$$\dot{\pi}_\beta[f] = \pi_\beta\left[f\dot{\pi}_\beta\right]$$

FISHER INFORMATION

FISHER INFORMATION

- The velocity $\dot{\pi}_\beta(x)$ measures the percentage change in mass at x from β to $\beta + \Delta\beta$

FISHER INFORMATION

- The velocity $\dot{\pi}_\beta(x)$ measures the percentage change in mass at x from β to $\beta + \Delta\beta$

$$\frac{\pi_{\beta'}(x)}{\pi_\beta(x)} = 1 + \Delta\beta\dot{\pi}_\beta(x) + O(\Delta\beta^2)$$

FISHER INFORMATION

- The velocity $\dot{\pi}_\beta(x)$ measures the percentage change in mass at x from β to $\beta + \Delta\beta$

$$\frac{\pi_{\beta'}(x)}{\pi_\beta(x)} = 1 + \Delta\beta\dot{\pi}_\beta(x) + O(\Delta\beta^2)$$

- Noteably when $f = 1$ we have

FISHER INFORMATION

- The velocity $\dot{\pi}_\beta(x)$ measures the percentage change in mass at x from β to $\beta + \Delta\beta$

$$\frac{\pi_{\beta'}(x)}{\pi_\beta(x)} = 1 + \Delta\beta\dot{\pi}_\beta(x) + O(\Delta\beta^2)$$

- Noteably when $f = 1$ we have

$$\pi_\beta[\dot{\pi}_\beta] = \frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[1] = \frac{\mathrm{d}}{\mathrm{d}\beta}1 = 0$$

FISHER INFORMATION

- The velocity $\dot{\pi}_\beta(x)$ measures the percentage change in mass at x from β to $\beta + \Delta\beta$

$$\frac{\pi_{\beta'}(x)}{\pi_\beta(x)} = 1 + \Delta\beta\dot{\pi}_\beta(x) + O(\Delta\beta^2)$$

- Noteably when $f = 1$ we have

$$\pi_\beta[\dot{\pi}_\beta] = \frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[1] = \frac{\mathrm{d}}{\mathrm{d}\beta}1 = 0$$

- The percentage change over $\mathbb{X}$ is zero as $\Delta\beta \to 0$

FISHER INFORMATION

- The velocity $\dot{\pi}_\beta(x)$ measures the percentage change in mass at x from β to $\beta + \Delta\beta$

$$\frac{\pi_{\beta'}(x)}{\pi_\beta(x)} = 1 + \Delta\beta\dot{\pi}_\beta(x) + O(\Delta\beta^2)$$

- Noteably when $f = 1$ we have

$$\pi_\beta[\dot{\pi}_\beta] = \frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[1] = \frac{\mathrm{d}}{\mathrm{d}\beta}1 = 0$$

  - The percentage change over $\mathbb{X}$ is zero as $\Delta\beta \to 0$

- We can measure the magnitude of the change using the variance deviation of $\dot{\pi}_\beta$ at β

FISHER INFORMATION

- The velocity $\dot{\pi}_\beta(x)$ measures the percentage change in mass at x from β to $\beta + \Delta\beta$

$$\frac{\pi_{\beta'}(x)}{\pi_\beta(x)} = 1 + \Delta\beta\dot{\pi}_\beta(x) + O(\Delta\beta^2)$$

- Noteably when $f = 1$ we have

$$\pi_\beta[\dot{\pi}_\beta] = \frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[1] = \frac{\mathrm{d}}{\mathrm{d}\beta}1 = 0$$

  - The percentage change over $\mathbb{X}$ is zero as $\Delta\beta \to 0$

- We can measure the magnitude of the change using the variance deviation of $\dot{\pi}_\beta$ at β
  - This coincides with **Fisher information** as the variance at β

FISHER INFORMATION

- The velocity $\dot{\pi}_\beta(x)$ measures the percentage change in mass at x from β to $\beta + \Delta\beta$

$$\frac{\pi_{\beta'}(x)}{\pi_\beta(x)} = 1 + \Delta\beta\dot{\pi}_\beta(x) + O(\Delta\beta^2)$$

- Noteably when $f = 1$ we have

$$\pi_\beta[\dot{\pi}_\beta] = \frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[1] = \frac{\mathrm{d}}{\mathrm{d}\beta}1 = 0$$

  - The percentage change over $\mathbb{X}$ is zero as $\Delta\beta \to 0$

- We can measure the magnitude of the change using the variance deviation of $\dot{\pi}_\beta$ at β
  - This coincides with **Fisher information** as the variance at β

$$I(\beta) := \mathbb{V}_\beta[\dot{\pi}_\beta] = \pi_\beta[\dot{\pi}_\beta^2]$$

FISHER INFORMATION

- The velocity $\dot{\pi}_\beta(x)$ measures the percentage change in mass at x from β to $\beta + \Delta\beta$

$$\frac{\pi_{\beta'}(x)}{\pi_\beta(x)} = 1 + \Delta\beta\dot{\pi}_\beta(x) + O(\Delta\beta^2)$$

- Noteably when $f = 1$ we have

$$\pi_\beta[\dot{\pi}_\beta] = \frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[1] = \frac{\mathrm{d}}{\mathrm{d}\beta}1 = 0$$

  - The percentage change over $\mathbb{X}$ is zero as $\Delta\beta \to 0$

- We can measure the magnitude of the change using the variance deviation of $\dot{\pi}_\beta$ at β
  - This coincides with **Fisher information** as the variance at β

$$I(\beta) := \mathbb{V}_\beta[\dot{\pi}_\beta] = \pi_\beta[\dot{\pi}_\beta^2]$$

  - Measures how sensitive the annealing path is to small changes in β

EXAMPLE GAUSSIAN PATHS

EXAMPLE GAUSSIAN PATHS

- **Example:** Suppose $\pi_\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ then the Fisther information equals

EXAMPLE GAUSSIAN PATHS

- **Example:** Suppose $\pi_\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ then the Fisther information equals

$$I(\beta) = \frac{\dot{\mu}_\beta^2 + 2\dot{\sigma}_\beta^2}{\sigma_\beta^2}$$

EXAMPLE GAUSSIAN PATHS

- **Example:** Suppose $\pi_\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ then the Fisther information equals

$$I(\beta) = \frac{\dot{\mu}_\beta^2 + 2\dot{\sigma}_\beta^2}{\sigma_\beta^2}$$

- If $\sigma_\beta = \sigma$ is constant then:

EXAMPLE GAUSSIAN PATHS

- **Example:** Suppose $\pi_\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ then the Fisther information equals

$$I(\beta) = \frac{\dot{\mu}_\beta^2 + 2\dot{\sigma}_\beta^2}{\sigma_\beta^2}$$

- If $\sigma_\beta = \sigma$ is constant then:

$$I(\beta) = \frac{\dot{\mu}_\beta^2}{\sigma^2}$$

EXAMPLE GAUSSIAN PATHS

- **Example:** Suppose $\pi_\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ then the Fisther information equals

$$I(\beta) = \frac{\dot{\mu}_\beta^2 + 2\dot{\sigma}_\beta^2}{\sigma_\beta^2}$$

- If $\sigma_\beta = \sigma$ is constant then:

$$I(\beta) = \frac{\dot{\mu}_\beta^2}{\sigma^2}$$

  - Changes in location correspond to linear changes in mass

EXAMPLE GAUSSIAN PATHS

- **Example:** Suppose $\pi_\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ then the Fisther information equals

$$I(\beta) = \frac{\dot{\mu}_\beta^2 + 2\dot{\sigma}_\beta^2}{\sigma_\beta^2}$$

- If $\sigma_\beta = \sigma$ is constant then:

$$I(\beta) = \frac{\dot{\mu}_\beta^2}{\sigma^2}$$

  - Changes in location correspond to linear changes in mass
  - The distribution changes more rapidly when σ is small for the unit change in μ

EXAMPLE GAUSSIAN PATHS

- **Example:** Suppose $\pi_\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ then the Fisther information equals

$$I(\beta) = \frac{\dot{\mu}_\beta^2 + 2\dot{\sigma}_\beta^2}{\sigma_\beta^2}$$

- If $\sigma_\beta = \sigma$ is constant then:

$$I(\beta) = \frac{\dot{\mu}_\beta^2}{\sigma^2}$$

  - Changes in location correspond to linear changes in mass
  - The distribution changes more rapidly when σ is small for the unit change in μ

- If $\mu_\beta = \mu$ is constant then:

EXAMPLE GAUSSIAN PATHS

- **Example:** Suppose $\pi_\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ then the Fisther information equals

$$I(\beta) = \frac{\dot{\mu}_\beta^2 + 2\dot{\sigma}_\beta^2}{\sigma_\beta^2}$$

- If $\sigma_\beta = \sigma$ is constant then:

$$I(\beta) = \frac{\dot{\mu}_\beta^2}{\sigma^2}$$

  - Changes in location correspond to linear changes in mass
  - The distribution changes more rapidly when σ is small for the unit change in μ

- If $\mu_\beta = \mu$ is constant then:

$$I(\beta) = \frac{\dot{\sigma}_\beta^2}{\sigma_\beta^2} = \left(\frac{\mathrm{d}}{\mathrm{d}\beta} \log \sigma_\beta \right)^2$$

EXAMPLE GAUSSIAN PATHS

- **Example:** Suppose $\pi_\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ then the Fisther information equals

$$I(\beta) = \frac{\dot{\mu}_\beta^2 + 2\dot{\sigma}_\beta^2}{\sigma_\beta^2}$$

- If $\sigma_\beta = \sigma$ is constant then:

$$I(\beta) = \frac{\dot{\mu}_\beta^2}{\sigma^2}$$

  - Changes in location correspond to linear changes in mass
  - The distribution changes more rapidly when σ is small for the unit change in μ

- If $\mu_\beta = \mu$ is constant then:

$$I(\beta) = \frac{\dot{\sigma}_\beta^2}{\sigma_\beta^2} = \left(\frac{\mathrm{d}}{\mathrm{d}\beta} \log \sigma_\beta \right)^2$$

  - Changes in scale correspond to logarithmic changes in mass independent of μ

CALCULUS OF ANNEALING PATHS

CALCULUS OF ANNEALING PATHS

- **Product rule:** If $f_\beta(x)$ is differentiable in β for all x with derivative $\dot{f}_\beta(x)$ with respect to β:

CALCULUS OF ANNEALING PATHS

- **Product rule:** If $f_\beta(x)$ is differentiable in β for all x with derivative $\dot{f}_\beta(x)$ with respect to β:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f_\beta] = \dot{\pi}_\beta[f_\beta] + \pi_\beta[\dot{f}_\beta]$$

CALCULUS OF ANNEALING PATHS

- **Product rule:** If $f_\beta(x)$ is differentiable in β for all x with derivative $\dot{f}_\beta(x)$ with respect to β:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f_\beta] = \dot{\pi}_\beta[f_\beta] + \pi_\beta[\dot{f}_\beta]$$

- The product rule is very powerful and allows us to do calculus on paths

CALCULUS OF ANNEALING PATHS

- **Product rule:** If $f_\beta(x)$ is differentiable in β for all x with derivative $\dot{f}_\beta(x)$ with respect to β:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f_\beta] = \dot{\pi}_\beta[f_\beta] + \pi_\beta[\dot{f}_\beta]$$

- The product rule is very powerful and allows us to do calculus on paths

- **Example:** take higher order derivatives in terms of the derivatives of f_β and $\dot{\pi}_\beta$. For exampl

CALCULUS OF ANNEALING PATHS

- **Product rule:** If $f_\beta(x)$ is differentiable in β for all x with derivative $\dot{f}_\beta(x)$ with respect to β:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f_\beta] = \dot{\pi}_\beta[f_\beta] + \pi_\beta[\dot{f}_\beta]$$

- The product rule is very powerful and allows us to do calculus on paths

- **Example:** take higher order derivatives in terms of the derivatives of f_β and $\dot{\pi}_\beta$. For exampl

- For example if f is contant in β,

$$\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\pi_\beta[f] = \pi_\beta[f(\dot{\pi}_\beta^2 + \ddot{\pi}_\beta)]$$

CALCULUS OF ANNEALING PATHS

- **Product rule:** If $f_\beta(x)$ is differentiable in β for all x with derivative $\dot{f}_\beta(x)$ with respect to β:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f_\beta] = \dot{\pi}_\beta[f_\beta] + \pi_\beta[\dot{f}_\beta]$$

- The product rule is very powerful and allows us to do calculus on paths

- **Example:** take higher order derivatives in terms of the derivatives of f_β and $\dot{\pi}_\beta$. For exampl
  - For example if f is contant in β,

$$\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\pi_\beta[f] = \pi_\beta[f(\dot{\pi}_\beta^2 + \ddot{\pi}_\beta)]$$

- **Example:** for any ϕ-divergence with twice-differentiable convex ϕ with $\phi(1) = \phi'(1) = 0$

CALCULUS OF ANNEALING PATHS

- **Product rule:** If $f_\beta(x)$ is differentiable in β for all x with derivative $\dot{f}_\beta(x)$ with respect to β:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f_\beta] = \dot{\pi}_\beta[f_\beta] + \pi_\beta[\dot{f}_\beta]$$

- The product rule is very powerful and allows us to do calculus on paths

- **Example:** take higher order derivatives in terms of the derivatives of f_β and $\dot{\pi}_\beta$. For exampl
  - For example if f is contant in β,

$$\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\pi_\beta[f] = \pi_\beta[f(\dot{\pi}_\beta^2 + \ddot{\pi}_\beta)]$$

- **Example:** for any ϕ-divergence with twice-differentiable convex ϕ with $\phi(1) = \phi'(1) = 0$

$$D_f(\pi_{\beta'} \| \pi_\beta) = \pi_\beta \left[\phi(1 + \Delta\pi_{\beta,\beta'})\right]$$

CALCULUS OF ANNEALING PATHS

- **Product rule:** If $f_\beta(x)$ is differentiable in β for all x with derivative $\dot{f}_\beta(x)$ with respect to β:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\pi_\beta[f_\beta] = \dot{\pi}_\beta[f_\beta] + \pi_\beta[\dot{f}_\beta]$$

- The product rule is very powerful and allows us to do calculus on paths

- **Example:** take higher order derivatives in terms of the derivatives of f_β and $\dot{\pi}_\beta$. For exampl
  - For example if f is contant in β,

$$\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\pi_\beta[f] = \pi_\beta[f(\dot{\pi}_\beta^2 + \ddot{\pi}_\beta)]$$

- **Example:** for any ϕ-divergence with twice-differentiable convex ϕ with $\phi(1) = \phi'(1) = 0$

$$D_f(\pi_{\beta'} \| \pi_\beta) = \pi_\beta \left[\phi(1 + \Delta\pi_{\beta,\beta'})\right]$$

$$= \frac{1}{2}\phi''(1)\Delta\beta^2 I(\beta) + O(\Delta\beta^3)$$

STEIN VS FISHER SCORE

STEIN VS FISHER SCORE

Fisher Score

STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

Stein Score

STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

Stein Score

$$\nabla \log \pi_\beta(x)$$

STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

Stein Score

$$\nabla \log \pi_\beta(x)$$

- Encodes the direction mass changes between distributions

$$\pi_{\beta'} \approx \pi_\beta + \Delta\beta \dot{\pi}_\beta$$

STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

- Encodes the direction mass changes between distributions

$$\pi_{\beta'} \approx \pi_\beta + \Delta\beta \dot{\pi}_\beta$$

Stein Score

$$\nabla \log \pi_\beta(x)$$

- Encodes the geometry of the modes

$$X_{t'} \approx X_t + \Delta t \nabla \log \pi_\beta(X_t)$$

STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

- Encodes the direction mass changes between distributions

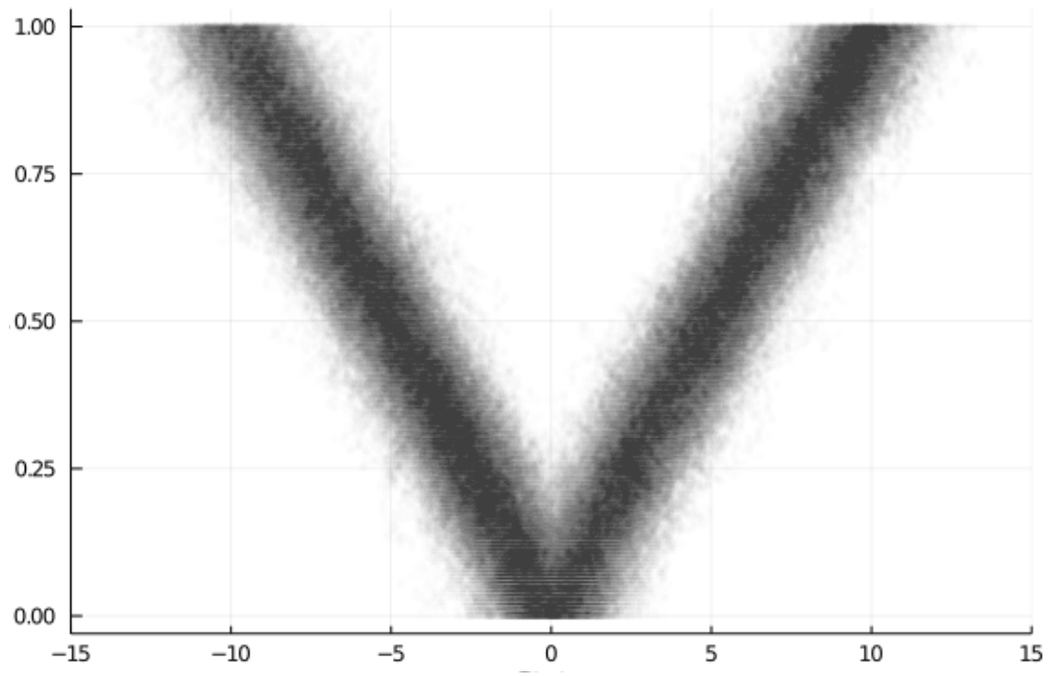
$$\pi_{\beta'} \approx \pi_\beta + \Delta\beta\dot{\pi}_\beta$$

Stein Score

$$\nabla \log \pi_\beta(x)$$

- Encodes the geometry of the modes

$$X_{t'} \approx X_t + \Delta t \nabla \log \pi_\beta(X_t)$$



STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\log \pi_\beta(x)$$

- Encodes the direction mass changes between distributions

$$\pi_{\beta'} \approx \pi_\beta + \Delta\beta\dot{\pi}_\beta$$

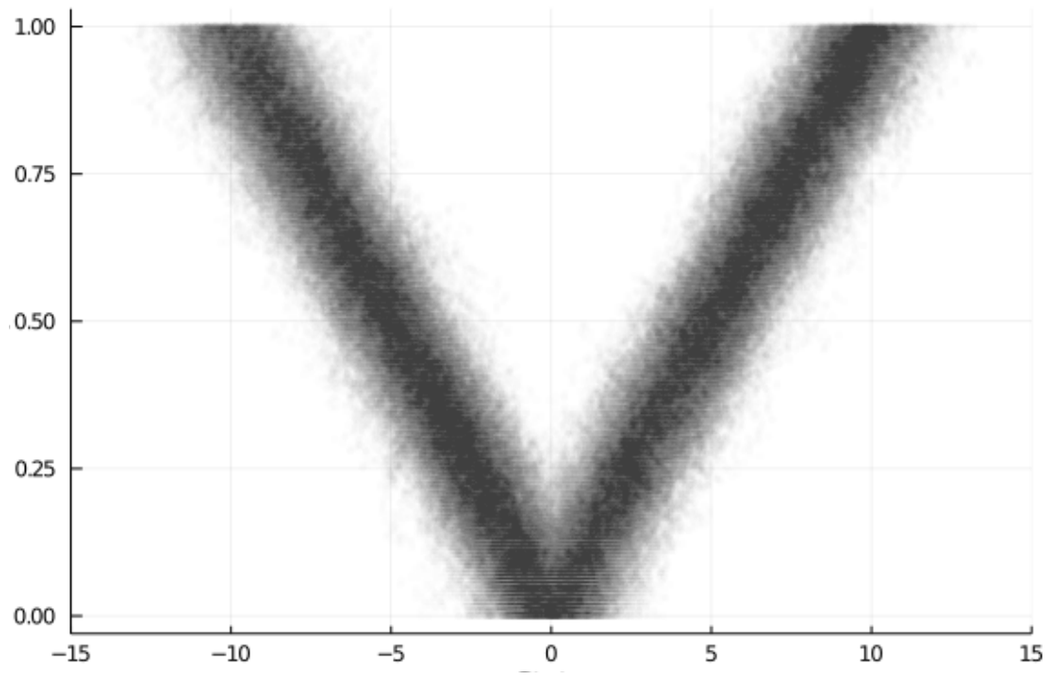
- Encodes the geometry of the normalising constant and modes

Stein Score

$$\nabla \log \pi_\beta(x)$$

- Encodes the geometry of the modes

$$X_{t'} \approx X_t + \Delta t \nabla \log \pi_\beta(X_t)$$



STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

- Encodes the direction mass changes between distributions

$$\pi_{\beta'} \approx \pi_\beta + \Delta\beta\dot{\pi}_\beta$$

- Encodes the geometry of the normalising constant and modes

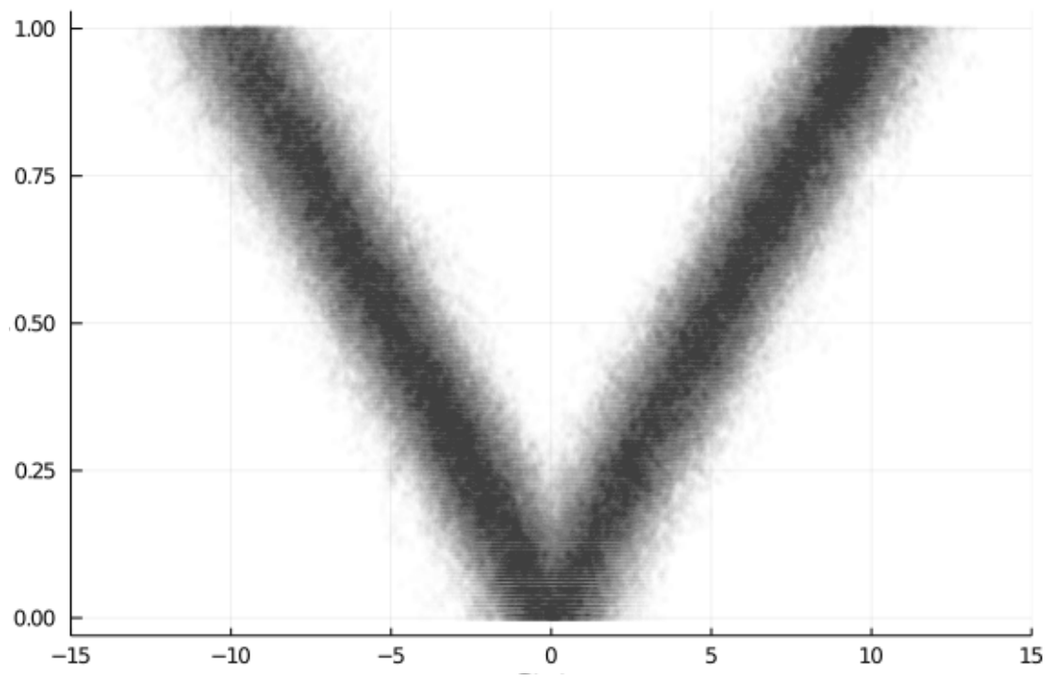
Stein Score

$$\nabla \log \pi_\beta(x)$$

- Encodes the geometry of the modes

$$X_{t'} \approx X_t + \Delta t \nabla \log \pi_\beta(X_t)$$

- Encodes the geometry of the modes



STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \pi_\beta(x)$$

- Encodes the direction mass changes between distributions

$$\pi_{\beta'} \approx \pi_\beta + \Delta\beta\dot{\pi}_\beta$$

- Encodes the geometry of the normalising constant and modes
- Controls for annealing

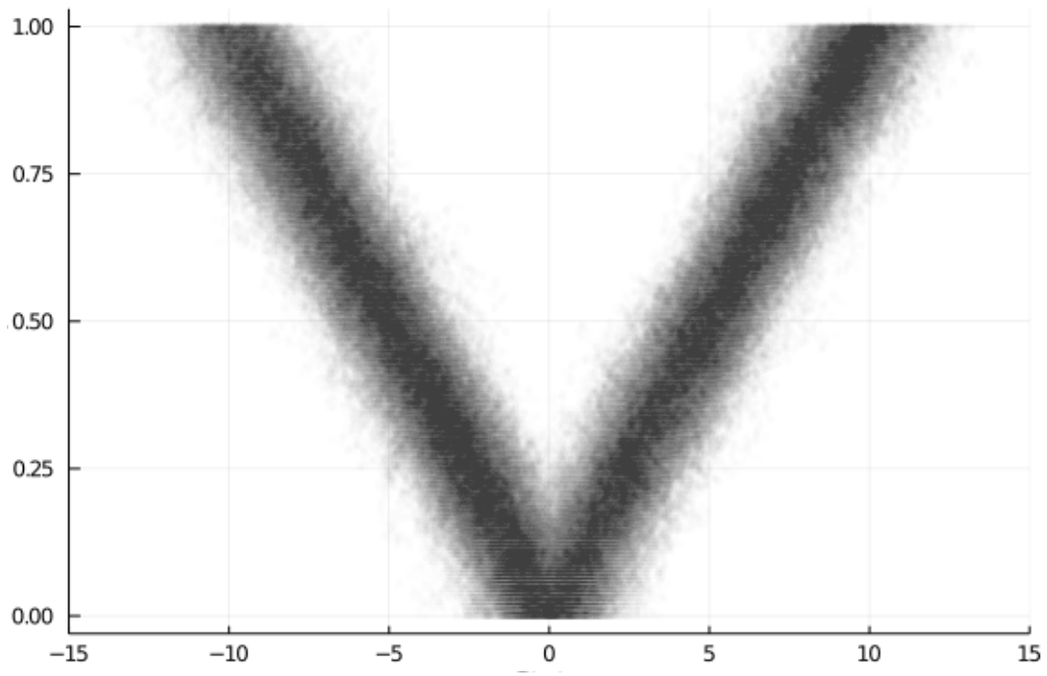
Stein Score

$$\nabla \log \pi_\beta(x)$$

- Encodes the geometry of the modes

$$X_{t'} \approx X_t + \Delta t \nabla \log \pi_\beta(X_t)$$

- Encodes the geometry of the modes



STEIN VS FISHER SCORE

Fisher Score

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\log \pi_\beta(x)$$

- Encodes the direction mass changes between distributions

$$\pi_{\beta'} \approx \pi_\beta + \Delta\beta\dot{\pi}_\beta$$

- Encodes the geometry of the normalising constant and modes
- Controls for annealing

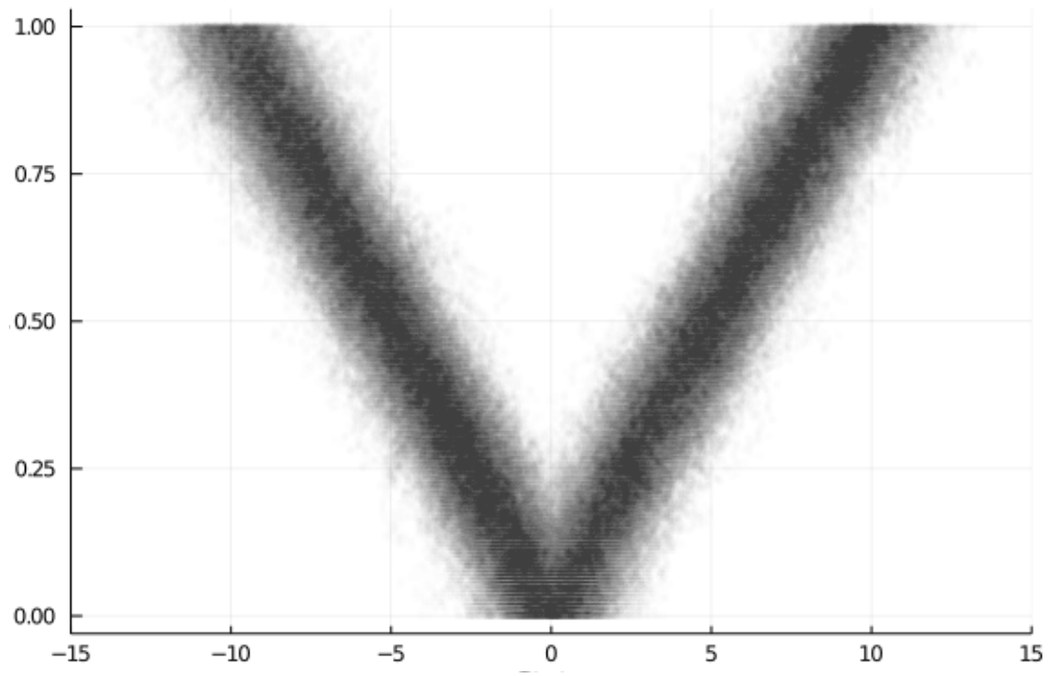
Stein Score

$$\nabla \log \pi_\beta(x)$$

- Encodes the geometry of the modes

$$X_{t'} \approx X_t + \Delta t \nabla \log \pi_\beta(X_t)$$

- Encodes the geometry of the modes
- Controls local inference



LINEAR PATH AS AN EXPONENTIAL FAMILY

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

  - $\Omega = \{\beta : Z(\beta) < \infty\}$ is an interval containing $[0,1]$

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

  - $\Omega = \{\beta : Z(\beta) < \infty\}$ is an interval containing $[0,1]$
  - η is the base-measure

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

  - $\Omega = \{\beta : Z(\beta) < \infty\}$ is an interval containing $[0,1]$
  - η is the base-measure
  - β is the natural parameter where

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

  - $\Omega = \{\beta : Z(\beta) < \infty\}$ is an interval containing $[0,1]$
  - η is the base-measure
  - β is the natural parameter where
  - $V = \log w$ is the sufficient statistic

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

  - $\Omega = \{\beta : Z(\beta) < \infty\}$ is an interval containing $[0,1]$
  - η is the base-measure
  - β is the natural parameter where
  - $V = \log w$ is the sufficient statistic
  - $A(\beta) = \log Z(\beta)$ is the cumulant generating function

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

  - $\Omega = \{\beta : Z(\beta) < \infty\}$ is an interval containing $[0,1]$
  - η is the base-measure
  - β is the natural parameter where
  - $V = \log w$ is the sufficient statistic
    - $A(\beta) = \log Z(\beta)$ is the cumulant generating function

- In particular, the smoothness of $A(\beta)$ controls the moments of V at π_β

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

  - $\Omega = \{\beta : Z(\beta) < \infty\}$ is an interval containing $[0,1]$
  - η is the base-measure
  - β is the natural parameter where
  - $V = \log w$ is the sufficient statistic
  - $A(\beta) = \log Z(\beta)$ is the cumulant generating function

- In particular, the smoothness of $A(\beta)$ controls the moments of V at π_β
  - The first derivative is the expectation of V with respect to π_β

$$A'(\beta) = \pi_\beta[V]$$

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

  - $\Omega = \{\beta : Z(\beta) < \infty\}$ is an interval containing $[0,1]$
  - η is the base-measure
  - β is the natural parameter where
  - $V = \log w$ is the sufficient statistic
  - $A(\beta) = \log Z(\beta)$ is the cumulant generating function

- In particular, the smoothness of $A(\beta)$ controls the moments of V at π_β
  - The first derivative is the expectation of V with respect to π_β

$$A'(\beta) = \pi_\beta[V]$$

  - The second derivative is the variance of V with respect to π_β i.e. the Fisher information

$$A''(\beta) = \mathbb{V}_\beta[V] = I(\beta)$$

LINEAR PATH AS AN EXPONENTIAL FAMILY

- The linear path extends naturally to an exponential family $\mathscr{E} = \{\pi_\beta : \beta \in \Omega\}$

$$\pi_\beta(x) = \exp(\beta V(x) - A(\beta))\eta(x)$$

  - $\Omega = \{\beta : Z(\beta) < \infty\}$ is an interval containing $[0,1]$
  - η is the base-measure
  - β is the natural parameter where
  - $V = \log w$ is the sufficient statistic
  - $A(\beta) = \log Z(\beta)$ is the cumulant generating function

- In particular, the smoothness of $A(\beta)$ controls the moments of V at π_β
  - The first derivative is the expectation of V with respect to π_β

$$A'(\beta) = \pi_\beta[V]$$

  - The second derivative is the variance of V with respect to π_β i.e. the Fisher information

$$A''(\beta) = \mathbb{V}_\beta[V] = I(\beta)$$

    - Hence $\beta \mapsto A(\beta)$ is convex with $A(0) = 0$

ANNEALING AS ROBUST BAYESIAN INFERNCE

ANNEALING AS ROBUST BAYESIAN INFERNCE

- Given a likelihood $L(y \mid x)$ and prior $p(x)$, the posterior $p(x \mid y)$

$$p(x \mid y) = \frac{L(x \mid y)p(x)}{p(y)}$$

ANNEALING AS ROBUST BAYESIAN INFERNCE

- Given a likelihood $L(y \mid x)$ and prior $p(x)$, the posterior $p(x \mid y)$

$$p(x \mid y) = \frac{L(x \mid y)p(x)}{p(y)}$$

- How sure are we that our model is correct? Are we robust to the likelihood, outliers, and the prior,

ANNEALING AS ROBUST BAYESIAN INFERNCE

- Given a likelihood $L(y \mid x)$ and prior $p(x)$, the posterior $p(x \mid y)$

$$p(x \mid y) = \frac{L(x \mid y)p(x)}{p(y)}$$

- How sure are we that our model is correct? Are we robust to the likelihood, outliers, and the prior,
- Are we robust to the likelihood, outliers, and the prior,

ANNEALING AS ROBUST BAYESIAN INFERNCE

- Given a likelihood $L(y \mid x)$ and prior $p(x)$, the posterior $p(x \mid y)$

$$p(x \mid y) = \frac{L(x \mid y)p(x)}{p(y)}$$

- How sure are we that our model is correct? Are we robust to the likelihood, outliers, and the prior,
- Are we robust to the likelihood, outliers, and the prior,
- We can measure the sensitivity of our inference by annealing the likelihood

$$p_\beta(x \mid y) = \frac{L_\beta(x \mid y)p(x)}{p_\beta(y)}$$

ANNEALING AS ROBUST BAYESIAN INFERNCE

- Given a likelihood $L(y \mid x)$ and prior $p(x)$, the posterior $p(x \mid y)$

$$p(x \mid y) = \frac{L(x \mid y)p(x)}{p(y)}$$

- How sure are we that our model is correct? Are we robust to the likelihood, outliers, and the prior,
- Are we robust to the likelihood, outliers, and the prior,
- We can measure the sensitivity of our inference by annealing the likelihood

$$p_\beta(x \mid y) = \frac{L_\beta(x \mid y)p(x)}{p_\beta(y)}$$

- Where L_β interpolates between 1 and L.

ANNEALING AS ROBUST BAYESIAN INFERNCE

- Given a likelihood $L(y \mid x)$ and prior $p(x)$, the posterior $p(x \mid y)$

$$p(x \mid y) = \frac{L(x \mid y)p(x)}{p(y)}$$

- How sure are we that our model is correct? Are we robust to the likelihood, outliers, and the prior,
- Are we robust to the likelihood, outliers, and the prior,
- We can measure the sensitivity of our inference by annealing the likelihood

$$p_\beta(x \mid y) = \frac{L_\beta(x \mid y)p(x)}{p_\beta(y)}$$

- Where L_β interpolates between 1 and L.
- Hence p_β interpolates between the prior and posterior

ANNEALING AS ROBUST BAYESIAN INFERNCE

- Given a likelihood $L(y \mid x)$ and prior $p(x)$, the posterior $p(x \mid y)$

$$p(x \mid y) = \frac{L(x \mid y)p(x)}{p(y)}$$

- How sure are we that our model is correct? Are we robust to the likelihood, outliers, and the prior,
- Are we robust to the likelihood, outliers, and the prior,
- We can measure the sensitivity of our inference by annealing the likelihood

$$p_\beta(x \mid y) = \frac{L_\beta(x \mid y)p(x)}{p_\beta(y)}$$

- Where L_β interpolates between 1 and L.
- Hence p_β interpolates between the prior and posterior
- Can see the influence of the data/model

LINEAR PATH AS POWER POSTERIOR

LINEAR PATH AS POWER POSTERIOR

- When $L_\beta(y \mid x) = L(y \mid x)^\beta$ we obtain the power posterior:

$$p_\beta(x \mid y) = \frac{L(x \mid y)^\beta p(x)}{p_\beta(y)}$$

LINEAR PATH AS POWER POSTERIOR

- When $L_\beta(y \mid x) = L(y \mid x)^\beta$ we obtain the power posterior:

$$p_\beta(x \mid y) = \frac{L(x \mid y)^\beta p(x)}{p_\beta(y)} = \frac{w^\beta(x)\eta(x)}{Z(\beta)} = \pi(x)$$

  - We can re-interpret the linear path in terms of Bayesian terminology:

LINEAR PATH AS POWER POSTERIOR

- When $L_\beta(y \mid x) = L(y \mid x)^\beta$ we obtain the power posterior:

$$p_\beta(x \mid y) = \frac{L(x \mid y)^\beta p(x)}{p_\beta(y)} = \frac{w^\beta(x)\eta(x)}{Z(\beta)} = \pi(x)$$

  - We can re-interpret the linear path in terms of Bayesian terminology:

- e.g. if the data is iid

LINEAR PATH AS POWER POSTERIOR

- When $L_\beta(y \mid x) = L(y \mid x)^\beta$ we obtain the power posterior:

$$p_\beta(x \mid y) = \frac{L(x \mid y)^\beta p(x)}{p_\beta(y)} = \frac{w^\beta(x)\eta(x)}{Z(\beta)} = \pi(x)$$

  - We can re-interpret the linear path in terms of Bayesian terminology:

- e.g. if the data is iid

$$L(y \mid x)^\beta = \prod_i L(y_i \mid x)^\beta$$

LINEAR PATH AS POWER POSTERIOR

- When $L_\beta(y \mid x) = L(y \mid x)^\beta$ we obtain the power posterior:

$$p_\beta(x \mid y) = \frac{L(x \mid y)^\beta p(x)}{p_\beta(y)} = \frac{w^\beta(x)\eta(x)}{Z(\beta)} = \pi(x)$$

  - We can re-interpret the linear path in terms of Bayesian terminology:

- e.g. if the data is iid

$$L(y \mid x)^\beta = \prod_i L(y_i \mid x)^\beta$$

  - $\beta < 1$ synthetically downweights data

LINEAR PATH AS POWER POSTERIOR

- When $L_\beta(y \mid x) = L(y \mid x)^\beta$ we obtain the power posterior:

$$p_\beta(x \mid y) = \frac{L(x \mid y)^\beta p(x)}{p_\beta(y)} = \frac{w^\beta(x)\eta(x)}{Z(\beta)} = \pi(x)$$

  - We can re-interpret the linear path in terms of Bayesian terminology:

- e.g. if the data is iid

$$L(y \mid x)^\beta = \prod_i L(y_i \mid x)^\beta$$

  - $\beta < 1$ synthetically downweights data
  - $\beta > 1$ synthetically identifying outliers in data

LINEAR PATH AS POWER POSTERIOR

- When $L_\beta(y \mid x) = L(y \mid x)^\beta$ we obtain the power posterior:

$$p_\beta(x \mid y) = \frac{L(x \mid y)^\beta p(x)}{p_\beta(y)} = \frac{w^\beta(x)\eta(x)}{Z(\beta)} = \pi(x)$$

  - We can re-interpret the linear path in terms of Bayesian terminology:

- e.g. if the data is iid

$$L(y \mid x)^\beta = \prod_i L(y_i \mid x)^\beta$$

  - $\beta < 1$ synthetically downweights data
  - $\beta > 1$ synthetically identifying outliers in data
  - Has applications in robust bayesian statistics, and coresets, etc

ANNEALING AS GLOBAL OPTIMISATION

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\arg\min_x f(x) + \lambda g(x)$$

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\arg\min_x f(x) + \lambda g(x) \quad = \arg\min_x \beta f(x) + g(x), \quad \beta = 1/\lambda$$

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\begin{aligned}\arg\min_x f(x) + \lambda g(x) \quad &= \arg\min_x \beta f(x) + g(x), \quad \beta = 1/\lambda \\ &= \arg\max_x \exp(-\beta f(x) - g(x))\end{aligned}$$

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\begin{aligned}\arg\min_x f(x) + \lambda g(x) &= \arg\min_x \beta f(x) + g(x), \quad \beta = 1/\lambda \\ &= \arg\max_x \exp(-\beta f(x) - g(x)) \\ &= \arg\max_x \pi_\beta(x)\end{aligned}$$

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\begin{aligned}\arg\min_x f(x) + \lambda g(x) &= \arg\min_x \beta f(x) + g(x), \quad \beta = 1/\lambda \\ &= \arg\max_x \exp(-\beta f(x) - g(x)) \\ &= \arg\max_x \pi_\beta(x)\end{aligned}$$

- Where for $\beta \geq 0$ define

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\begin{aligned}\arg\min_x f(x) + \lambda g(x) &= \arg\min_x \beta f(x) + g(x), \quad \beta = 1/\lambda \\ &= \arg\max_x \exp(-\beta f(x) - g(x)) \\ &= \arg\max \pi_\beta(x)\end{aligned}$$

- Where for $\beta \geq 0$ define

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\begin{aligned}\arg\min_x f(x) + \lambda g(x) &= \arg\min_x \beta f(x) + g(x), \quad \beta = 1/\lambda \\ &= \arg\max_x \exp(-\beta f(x) - g(x)) \\ &= \arg\max \pi_\beta(x)\end{aligned}$$

- Where for $\beta \geq 0$ define

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

- We can view the linear path as encoding MAP estimate for global optimization problem

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\begin{aligned}\arg\min_x f(x) + \lambda g(x) &= \arg\min_x \beta f(x) + g(x), \quad \beta = 1/\lambda \\ &= \arg\max_x \exp(-\beta f(x) - g(x)) \\ &= \arg\max \pi_\beta(x)\end{aligned}$$

- Where for $\beta \geq 0$ define

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

- We can view the linear path as encoding MAP estimate for global optimization problem
  - The regulariser is encoded by the reference

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\begin{aligned}\arg\min_x f(x) + \lambda g(x) &= \arg\min_x \beta f(x) + g(x), \quad \beta = 1/\lambda \\ &= \arg\max_x \exp(-\beta f(x) - g(x)) \\ &= \arg\max \pi_\beta(x)\end{aligned}$$

- Where for $\beta \geq 0$ define

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

- We can view the linear path as encoding MAP estimate for global optimization problem
  - The regulariser is encoded by the reference
  - The annealing parameter encores the level of refularisation

ANNEALING AS GLOBAL OPTIMISATION

- Suppose we wish to minimize $f(x)$ possibly non-convex function.
- Consider the regularised problem for $\lambda > 0$:

$$\begin{aligned}\arg\min_x f(x) + \lambda g(x) &= \arg\min_x \beta f(x) + g(x), \quad \beta = 1/\lambda \\ &= \arg\max_x \exp(-\beta f(x) - g(x)) \\ &= \arg\max \pi_\beta(x)\end{aligned}$$

- Where for $\beta \geq 0$ define

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

- We can view the linear path as encoding MAP estimate for global optimization problem
  - The regulariser is encoded by the reference
  - The annealing parameter encores the level of refularisation
  - Objective is encoded by the weight

ANNEALING AS GLOBAL OPTIMISATION

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

$$\propto \exp(-\beta U(x) - g(x)), \quad U(x) = f(x) - \min_x f(x)$$

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

$$\propto \exp(-\beta U(x) - g(x)), \quad U(x) = f(x) - \min_x f(x)$$

$$= 1[U(x) = 0] + 1[U(x) > 0]\exp(-\beta U(x) - g(x))$$

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

$$\propto \exp(-\beta U(x) - g(x)), \quad U(x) = f(x) - \min_x f(x)$$

$$= 1[U(x) = 0] + 1[U(x) > 0]\exp(-\beta U(x) - g(x))$$

  - As $\beta \to \infty$ we have the second term goes to zero and π_β concentrates on the null set of U

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\begin{aligned}\pi_\beta(x) &\propto \exp(-\beta f(x) - g(x)) \\ &\propto \exp(-\beta U(x) - g(x)), \quad U(x) = f(x) - \min_x f(x) \\ &= 1[U(x) = 0] + 1[U(x) > 0]\exp(-\beta U(x) - g(x))\end{aligned}$$

  - As $\beta \to \infty$ we have the second term goes to zero and π_β concentrates on the null set of U

$$\lim_{\beta \to \infty} \pi_\beta = \delta_{U=0}$$

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\begin{aligned}\pi_\beta(x) &\propto \exp(-\beta f(x) - g(x)) \\ &\propto \exp(-\beta U(x) - g(x)), \quad U(x) = f(x) - \min_x f(x) \\ &= 1[U(x) = 0] + 1[U(x) > 0]\exp(-\beta U(x) - g(x))\end{aligned}$$

  - As $\beta \to \infty$ we have the second term goes to zero and π_β concentrates on the null set of U

$$\lim_{\beta \to \infty} \pi_\beta = \delta_{U=0}$$

- Note that $U(x) = 0$ is precicely when $f(x)$ is minimised

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\begin{aligned}\pi_\beta(x) &\propto \exp(-\beta f(x) - g(x)) \\ &\propto \exp(-\beta U(x) - g(x)), \quad U(x) = f(x) - \min_x f(x) \\ &= 1[U(x) = 0] + 1[U(x) > 0]\exp(-\beta U(x) - g(x))\end{aligned}$$

  - As $\beta \to \infty$ we have the second term goes to zero and π_β concentrates on the null set of U

$$\lim_{\beta \to \infty} \pi_\beta = \delta_{U=0}$$

- Note that $U(x) = 0$ is precicely when $f(x)$ is minimised

$$U(x) = 0 \quad \iff \quad x \in \arg\min_x f(x)$$

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\begin{aligned}\pi_\beta(x) &\propto \exp(-\beta f(x) - g(x)) \\ &\propto \exp(-\beta U(x) - g(x)), \quad U(x) = f(x) - \min_x f(x) \\ &= 1[U(x) = 0] + 1[U(x) > 0]\exp(-\beta U(x) - g(x))\end{aligned}$$

  - As $\beta \to \infty$ we have the second term goes to zero and π_β concentrates on the null set of U

$$\lim_{\beta \to \infty} \pi_\beta = \delta_{U=0}$$

  - Note that $U(x) = 0$ is precicely when $f(x)$ is minimised

$$U(x) = 0 \quad \iff \quad x \in \arg\min_x f(x)$$

- Sampling from π_β for large β is equivalent to global optimising

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\begin{aligned}\pi_\beta(x) &\propto \exp(-\beta f(x) - g(x)) \\ &\propto \exp(-\beta U(x) - g(x)), \quad U(x) = f(x) - \min_x f(x) \\ &= 1[U(x) = 0] + 1[U(x) > 0]\exp(-\beta U(x) - g(x))\end{aligned}$$

  - As $\beta \to \infty$ we have the second term goes to zero and π_β concentrates on the null set of U

$$\lim_{\beta \to \infty} \pi_\beta = \delta_{U=0}$$

  - Note that $U(x) = 0$ is precicely when $f(x)$ is minimised

$$U(x) = 0 \quad \iff \quad x \in \arg\min_x f(x)$$

- Sampling from π_β for large β is equivalent to global optimising
- Note that we did not make an assumption on the state-space

ANNEALING AS GLOBAL OPTIMISATION

- To see what happens as $\beta \to \infty$, note that we can write π_β as

$$\pi_\beta(x) \propto \exp(-\beta f(x) - g(x))$$

$$\propto \exp(-\beta U(x) - g(x)), \quad U(x) = f(x) - \min_x f(x)$$

$$= 1[U(x) = 0] + 1[U(x) > 0]\exp(-\beta U(x) - g(x))$$

  - As $\beta \to \infty$ we have the second term goes to zero and π_β concentrates on the null set of U

$$\lim_{\beta \to \infty} \pi_\beta = \delta_{U=0}$$

  - Note that $U(x) = 0$ is precicely when $f(x)$ is minimised

$$U(x) = 0 \quad \iff \quad x \in \arg\min_x f(x)$$

- Sampling from π_β for large β is equivalent to global optimising
- Note that we did not make an assumption on the state-space
- Annealing methods were first introduced to tackle hard combinatorial optimization problems